UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE POS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Fábio Bif Goularte

MÉTODO FUZZY PARA A SUMARIZAÇÃO AUTOMÁTICA DE TEXTO COM BASE EM UM MODELO EXTRATIVO (FSumm)

Florianópolis
2015

Fábio Bif Goularte

MÉTODO FUZZY PARA A SUMARIZAÇÃO AUTOMÁTICA DE TEXTO COM BASE EM UM MODELO EXTRATIVO (FSumm)

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciência da Computação.
Orientadora: Prof.ª Dr.ª Silvia Modesto Nassar

Florianópolis
2015

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Goularte, Fábio Bif
Método Fuzzy para a Sumarização Automática de Texto com base em um Modelo Extrativo (FSumm) / Fábio Bif Goularte ; orientador, Silvia Modesto Nassar - Florianópolis, SC, 2015.
117 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação.

Inclui referências

1. Ciência da Computação. 2. Sumarização automática de texto. 3. Lógica fuzzy. 4. Métricas de extração de informação. I. Nassar, Silvia Modesto. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Fábio Bif Goularte

MÉTODO FUZZY PARA A SUMARIZAÇÃO AUTOMÁTICA DE TEXTO COM BASE EM UM MODELO EXTRATIVO (FSUMM)

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Ciência da Computação, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Florianópolis, 23 de fevereiro de 2015.

_____
Prof. Ronaldo Mello, Dr.
Coordenador do programa

_____
Prof.ª Silvia Modesto Nassar, Dr.ª
Orientadora

Banca Examinadora:

_____
Prof.ª Aline Villavicencio, Dr.ª
Universidade Federal do Rio Grande do Sul

_____
Prof. Mauro Roinsenberg, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Renato Fileto, Dr.
Universidade Federal de Santa Catarina

Esta dissertação é dedicada a todos àqueles que de forma direta ou indireta contribuíram para que fosse possível.

AGRADECIMENTOS

À minha orientadora, Silvia, primeiramente, por conceder a oportunidade e me acompanhar nesta dissertação, de confiar no meu modo de trabalho e acreditar, junto comigo, que era possível.

Aos professores Aline Villavicencio, Mauro Roisenberg e Renato Fileto por aceitarem estar na banca de defesa e contribuir com esta dissertação.

Aos colegas do laboratório e de projeto SAAS e-Tec, e aos professores: Renato Cislaghi, pelo auxílio e trabalho no projeto SAAS; Masanao Ohirama pela disponibilidade de infraestrutura para trabalho, assim como a UFSC.

Aos meus pais, Maria e Valdemar, por sempre me aconselharem e ajudar nos momentos difíceis. Procuro não decepcioná-los.

Aos meus irmãos, Edena, Dalmir e Dilma, e aos meus sobrinhos Michele, Manolla, Samuel, Nabilla e Tamara pelos incontáveis momentos de sorriso largo e conversas alegres a respeito de qualquer assunto. Hoje somos adultos e amigos, o que me deixa muito feliz! ☺

Aos amigos que fiz pelos lugares em que passei, em especial ao Jonas por abrir meus olhos e inspirar-me a buscar aquilo que eu mais queria.

Ao pessoal do projeto amanhecer do HU da UFSC por cuidarem de mim em momentos difíceis.

A Deus, pela vida e oportunidades que nela existem.

RESUMO

A sumarização automática de texto procura condensar o conteúdo do documento, extraindo as informações mais relevantes. Esse processo normalmente é executado através de métodos computacionais que incorporam o método estatístico e o linguístico. O rápido desenvolvimento das tecnologias emergentes e a crescente quantidade de informação disponível inserem novos desafios para esta área de pesquisa. Um desses desafios está na identificação das sentenças mais informativas no momento da geração do sumário. Como a tarefa de sumarizar informações de texto traz consigo a incerteza inerente à linguagem natural, a lógica *fuzzy* pode ser aplicada nessa tarefa para contribuir nos resultados gerados. Portanto, esta dissertação propõe um método de sumarização automática de texto utilizando a lógica *fuzzy* para a classificação das sentenças. O método foi desenvolvido por meio da técnica de sumarização extrativa ao qual se associam tarefas de Recuperação de Informação (RI) e de Processamento de Linguagem Natural (PLN). Para a avaliação deste método, considerou-se um corpus de textos em língua portuguesa e uma ferramenta que automatiza o processo. A ferramenta de avaliação analisa a sobreposição das unidades textuais entre os sumários automáticos e o modelo humano, dadas pelas medidas de precisão, cobertura e medida-f. Foram realizados experimentos que demonstram a efetividade do método na classificação da informatividade das sentenças.

Palavras-chave: Sumarização automática de texto. Lógica *fuzzy*. Métricas de extração de informação.

ABSTRACT

Automatic text summarization attempts to condense the document content, extracting the most relevant information. This process is usually performed by computational methods such as statistical and linguistic. The rapid development of emerging technologies and the increasing amount of information available insert new research challenges. One of these challenges is to identify the most informative sentences at the time of the summary generation. The textual information summarization task brings with it the uncertainty inherent in natural language where fuzzy logic can be applied and contribute to the results. Therefore, this dissertation proposes a method of automatic text summarization using fuzzy logic to the classification of sentences. The method was developed by extractive summarization techniques which are associated with information retrieval tasks (IR) and natural language processing (NLP). The evaluation method considers a corpus of Brazilian Portuguese news texts and a tool for evaluation of summaries. The assessment tool analyzes the text units overlaps between automatic summaries and human model producing measures (precision, recall, F-measure) that express the informativeness of the summaries. We also present experiments showing the effectiveness of our method in the informativeness sentences classification.

Palavras-chave: Automatic text summarization. Fuzzy logic. Information extraction metrics.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AutoSummENG – *AUTOmatic SUMMary Evaluation based on N-gram Graphs*
BE – *Basic Elements*
BEwT-E – *Basic Elements with Transformations for Evaluation*
CST – *Cross-document Structure Theory*
DUC – *Document Understanding Conference*
GEMS – *Generative Modeling for Evaluation of Summaries*
IC – Inteligência Computacional
len – comprimento
loc – posição
MO – monodocumento
MQO – Mínimos Quadrados Ordinários
MVC – Model-*View-Controller*
PLN – Processamento de Linguagem Natural
POS – *tags parts-of-speech*
PROPOR – *International Conference on Computational Processing of Portuguese*
RI – Recuperação da Informação
ROUGE – *Recall-Oriented Understudy for Gisting Evaluation*
SAT – Sumarização Automática de Texto
SIF – Sistema de Inferência Fuzzy
SLF – Sistema de Lógica Fuzzy
ST – Sumarização de Texto
TAC – *Text Analysis Conference*
tf-idf – *Term Frequency - Inverse Document Frequency*
tf-isf – *Term Frequency - Inverse Sentence Frequency*
TKS – Takagi-Sugeno-Kang
TREC – *Text Retrieval Conference*
UCS – Unidades de conteúdo semântico
UM – multidocumento

SUMÁRIO

1 INTRODUÇÃO

Sumarizar é resumir um texto ou conjunto de textos, extraindo as informações mais importantes sem que este perca o sentido do texto original (NENKOVA; MCKEOWN, 2011). A sumarização automática de texto caracteriza-se pela geração de um sumário através de métodos computacionais associados a técnicas de Recuperação da Informação e Mineração de Texto (AGGARWAL; ZHAI, 2012).

A sumarização é um processo inerente a mente humana e a sumarização automática de texto busca automatizar o processo humano de produção de sumários.

A informação na forma textual é representada por termos linguísticos que trazem consigo a incerteza. A incerteza envolvida na resolução de um problema pode ser decorrente de alguma informação deficiente ou porque existe mais de uma solução. Para a sumarização de texto, essa deficiência pode ser originada de forma cognitiva – inerente à ambiguidade da linguagem natural.

A modelagem *fuzzy* é uma importante abordagem amplamente recomendada para aplicações cujo domínio esteja caracterizado por incerteza ou imprecisão da informação. A lógica *fuzzy* é capaz de combinar a imprecisão associada aos eventos naturais e o poder computacional de máquinas para produzir sistemas inteligentes (DAS, 2013).

Segundo Ross (2010), a lógica *fuzzy* se aproxima da forma com que o raciocínio humano, diante da incerteza, relaciona as informações buscando respostas aproximadas aos problemas. Além disso, tem-se uma ferramenta capaz de representar os termos utilizados na linguagem natural.

A ampla utilização da web e de tecnologias, tais como: dispositivos móveis, as redes sociais, os sistemas de informação eletrônicos, têm ocasionado um aumento exponencial do volume de informações e dados que são publicados. Segundo Hilbert e López (2011) a quantidade de dados que estão sendo coletados ultrapassou a capacidade humana de armazená-los. O mundo de hoje é baseado na informação, sendo que a maior parte está online (NENKOVA; MCKEOWN, 2011).

No caso da web, a crescente quantidade de informação disponível dificulta identificar o que é relevante, e requer tempo de pesquisa (GUPTA; LEHAL, 2010; LLORET; PALOMAR, 2012). Basta realizar uma pesquisa em um motor de busca que se verá a massiva quantidade

de páginas retornadas e os diferentes formatos disponíveis da informação (texto, imagem, som, vídeo).

A sobrecarga de informação pode, ainda, não ser vista como um problema, porém, a forma como o ser humano a manipula e consome deve ser repensada. Sendo assim, há necessidade de novas abordagens para organização da informação frente ao cenário apresentado, e esta preocupação têm suscitado o interesse científico no desenvolvimento de sistemas de sumarização automática.

1.1 MOTIVAÇÃO

Atualmente, sistemas de notícias baseados na Web, por exemplo, o NewsBlaster, são capazes de identificar e lidar com a redundância de documentos, o que assegura sumários completos e coerentes (BARZILAY; MCKEOWN, 2005).

As principais propriedades para um bom sumário são: relevância – o sumário deve conter unidades textuais informativas e relevantes para o usuário; redundância – não devem conter várias unidades textuais que transmitam a mesma informação; comprimento – tamanho limitado.

Aperfeiçoar essas três propriedades em conjunto é algo desafiador, é um exemplo de problema de compactação global[1]. A inclusão de unidades textuais relevantes depende não apenas das propriedades das unidades em si, mas também das propriedades de qualquer outra unidade textual no todo.

Um texto sumarizado pode contribuir para reduzir os esforços de pesquisa e leitura de usuários (pessoa ou agente) na Web, assim como em contextos onde seja necessário lidar com um grande volume de informação textual, por exemplo: na área da educação, processos judiciais, laudos médicos, relatórios econômicos.

No caso da área da educação, os estudantes e professores estão, cada vez mais, familiarizados com *tablets*. Como o dispositivo é utilizado à leitura de livro digital, material didático e pesquisa; a sumarização de texto pode ser uma forma efetiva de contribuir para a potencialização do processo ensino-aprendizagem (YANG et al., 2013; GOULARTE; WILGES; NASSAR, 2014).

[1] Na Teoria da Complexidade Computacional, o problema de compactação global pode ser classificado em NP-completo (ALGULIEV; ALYGULIEV, 2008). A geração de um algoritmo que encontre a solução ótima para esse problema é computacionalmente inviável, porém é possível desenvolver uma solução que se aproxima da solução ótima.

1.2 PROBLEMATIZAÇÃO

Grande parte da informação está na forma textual e devido ao volume, identificar o que é relevante e de forma coerente remetem a capacidade de sumarizar. O fluxo de informações em um documento não é uniforme, o que significa que algumas partes são mais importantes do que outras (DAS; MARTINS, 2007). O grande desafio na sumarização encontra-se em distinguir quais são as partes mais informativas.

Um texto apresenta propriedades ou características que são mensuradas por métricas. Quais são as propriedades textuais que devem ser retidas no sumário? Quais as melhores métricas para alcançar o conteúdo de documentos? Essas perguntas são subjetivas e difíceis de responder, pois a pontuação das características traz consigo a incerteza e a imprecisão.

Neste caso, imaginar que pode haver uma caraterística que seja mais importante que outra é algo plausível. Uma das primeiras tentativas de lidar com a importância das características foi a abordagem proposta por Edmundson (1969), que define a importância de uma característica por meio de parâmetros ajustáveis. Edmundson também foi o responsável pela ideia de que um conjunto de características textuais, e não apenas uma, pode determinar o conteúdo importante de um texto.

Desde então, abordagens de sumarização de texto que consideram um conjunto de características textuais têm sido propostas apoiadas em algoritmos Bayesianos, lógica *fuzzy*, Modelos de Cadeias de Markov, Redes Neurais, Modelos de Regressão, algoritmos gulosos, grafos, entre outros, cada qual com vantagens e desvantagens.

Quando a abordagem é baseada em regras, uma das desvantagens é a baixa capacidade de adaptação das regras para novos domínios, mas tem a vantagem, em geral, de apresentar desempenho melhor que outras abordagens. Já com as abordagens baseadas em aprendizado de máquina, a desvantagem é a necessidade de grandes quantidades de dados de treinamentos, mas tem a vantagem da independência de domínio (NENKOVA; MCKEOWN, 2011).

Ao observarem-se os métodos de sumarização de texto encontrados na literatura, em especifico os que aplicam a lógica *fuzzy* (ver o estado da arte na seção 2.3), fica evidente que são utilizadas várias características textuais e, da mesma forma, várias métricas (pelo menos uma métrica para cada característica).

Contudo, independente da abordagem, a quantidade de características textuais e métricas podem ser cruciais para o desempenho

do método de sumarização, pois à medida que aumenta a quantidade de características e métricas, também aumenta a complexidade e a dimensão do problema.

Com o intuito de explorar um formalismo para representar a incerteza por imprecisão na extração de informação, a principal contribuição desta pesquisa é um novo método *fuzzy* de sumarização de texto que manipula um conjunto de métricas extrativas através da correlação de características textuais. Por meio da correlação das características textuais é possível diminuir a quantidade de métricas.

Dessa forma, a questão a ser investigada é a de que o conteúdo informativo de textos pode ser identificado pelo tratamento impreciso de características correlacionadas.

A sumarização de texto é um processo inerente à mente humana, como já foi comentado, e *fuzzy* é uma forma de modelar a lógica do raciocínio humano, mais condizente com a realidade. A lógica *fuzzy* aplicada a sumarização de texto, por meio de funções de pertinência, é capaz de lidar com as diferentes características do texto, sem a necessidade de otimização de parâmetros para atribuir a importância das características.

O presente trabalho se enquadra na área de Ciência da Computação, na linha de pesquisa Inteligência Computacional (IC), especificamente em Processamento de Linguagem Natural (*Natural Language Processing* - PLN). Um dos objetivos do PLN consiste em desenvolver pesquisas na área de reconhecimento e produção de informações apresentadas em linguagem natural.

1.3 OBJETIVOS

O objetivo geral desta pesquisa é propor um sistema de inferência fuzzy composto por um conjunto de métricas extrativas que estimam a informatividade de uma sentença na sumarização automática de texto.

Vinculado ao objetivo geral e ao problema investigado, os objetivos específicos são:

- Identificar as características textuais importantes para o processo de sumarização automática de texto pela técnica extrativa;
- Identificar com base na literatura as características mais importantes para compor o método *fuzzy* de sumarização;

- Propor métricas de sumarização por meio da análise das propriedades do corpus TeMário.
- Avaliar o desempenho do método proposto em relação a informatividade dos sumários.

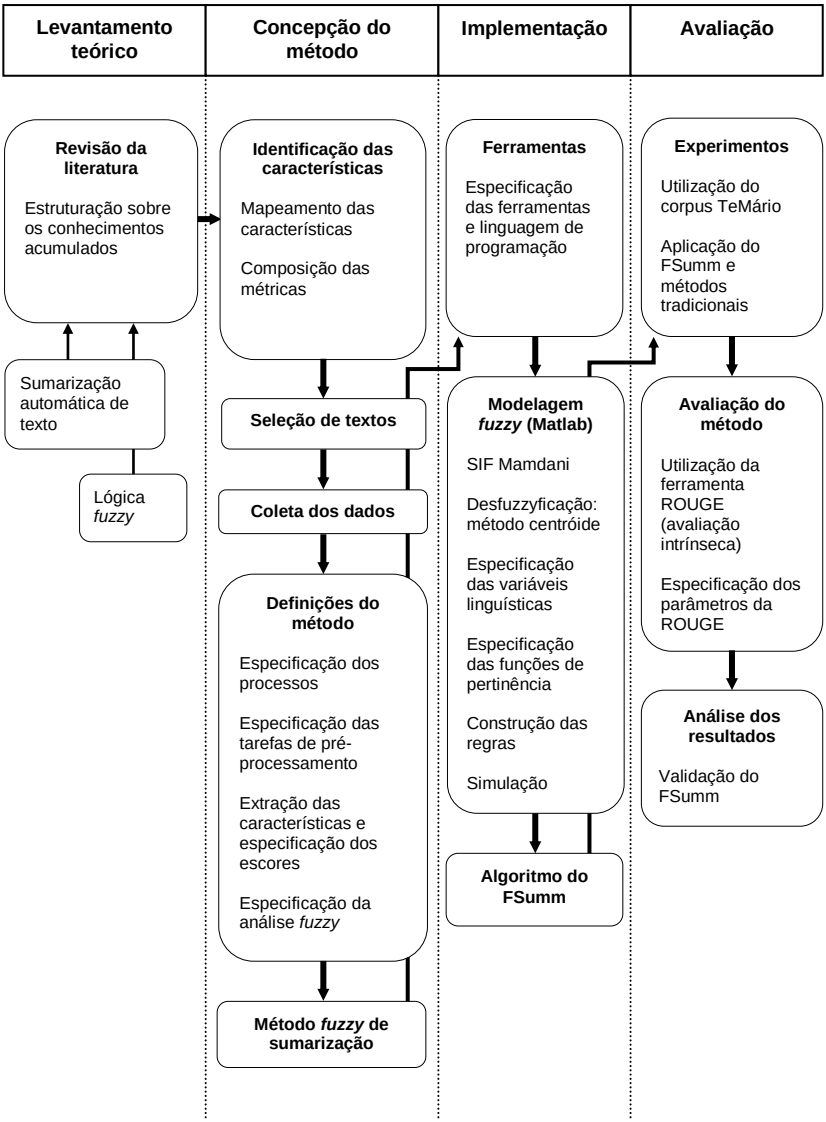
1.4 METODOLOGIA

Esta dissertação planeja e desenvolve um método *fuzzy* de sumarização de texto por meio da análise de características do texto. Nesse sentido, estabelecer o desenho ou modelo, o tipo e os procedimentos metodológicos da pesquisa mostram uma sequência lógica que deve ser seguida para alcançar os objetivos do estudo.

Quanto ao tipo da pesquisa, esse estudo é de natureza aplicada e de base tecnológica, pois visa gerar conhecimentos dirigidos a solução de um problema específico através do empirismo e técnicas de IC.

Os procedimentos metodológicos previstos na construção do método são apresentados na Figura 1 que, em síntese, correspondem ao desenho da pesquisa. Além do levantamento teórico é necessário conceber, implementar e avaliar o método proposto.

Figura 1 – Desenho da pesquisa



Fonte: Elaborada pelo autor

A primeira fase da pesquisa é a revisão de literatura, quando são levantados os artigos e obras disponíveis em bibliotecas digitais e base de dados na internet, sobre os temas que compõem a fundamentação teórica, especificamente sobre sumarização automática de texto e lógica *fuzzy*.

A fase seguinte refere-se à especificação e a descrição do método proposto, onde é necessário identificar as características textuais e eleger, com base na literatura, as que serão usadas na composição das métricas. Também é realizada, nessa fase, a seleção de um conjunto de textos, a coleta dos dados e a definição do método. A definição do método está compreendida em três processos que são: pré-processamento, análise das características e métricas e análise *fuzzy*, cada qual subdividido em etapas pontuais para então conceber o método *fuzzy* de sumarização.

A implementação é a fase onde uma aplicação do método proposto é desenvolvida. A aplicação integra os três processos mencionados na fase anterior. A modelagem *fuzzy* que compreende a definição do sistema de inferência *fuzzy* (SIF), o método de desfuzzyficação, a especificação das variáveis linguísticas, das funções de pertinências, a construção das regras e a simulação do sistema é realizada em uma ferramenta específica e posteriormente integrada à aplicação.

Com a implementação do FSumm é possível acessar a fase de avaliação, que assim como a fase de implementação são consideradas as partes práticas deste estudo. Na fase de avaliação definiram-se os experimentos de avaliação do FSumm. A ferramenta de avaliação automatizada (ROUGE) que compara os sumários de referência, produzidos por especialista humano, com os sumários produzidos pelo FSumm foi utilizada e apresentado os resultados.

1.5 ESTRUTURA DA DISSERTAÇÃO

Os próximos capítulos estão organizados da seguinte forma: no Capítulo 2 são encontrados os principais conceitos sobre os temas que fazem parte da fundamentação teórica. Uma revisão de literatura envolvendo a sumarização de texto extrativa e a lógica *fuzzy* é apresentada, e finaliza descrevendo os principais trabalhos publicados na área em que esta dissertação está inserida. No Capítulo 3 é apresentada a proposta de geração de sumários, com os procedimentos metodológicos, as etapas necessárias para o planejamento e construção do método, a implementação do método e as propriedades matemáticas

definidas pelas métricas. No Capítulo 4, apresenta-se a avaliação, a qual faz parte à seleção de um corpus, a definição dos experimentos e os resultados com a discussão. Por fim, são apresentadas as considerações, as contribuições e algumas perspectivas para a continuação deste trabalho no Capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

Nessa análise abordam-se as definições de sumarização automática de texto, as principais abordagens e métodos de sumarização extrativa, as características do texto, as métricas, as formas de avaliação e a classificação com lógica *fuzzy*.

2.1 A SUMARIZAÇÃO DE TEXTO

Quando uma pessoa relata ou escreve sobre um texto que leu, ela está produzindo uma versão mais curta do texto com as informações que julga importante, em outras palavras, um sumário.

A Sumarização de Texto (*Text Summarization* - ST) foi definida por Jones (1999) como sendo uma transformação redutiva de um texto fonte para um texto resumido ou sumário, através da seleção e/ou generalização do conteúdo importante do texto-fonte. Quando essa redução de texto é realizada por um computador, recebe o nome de Sumarização Automática de Texto (*Automatic Text Summarization* - SAT).

A SAT é o processo onde um ou mais documentos servem de entrada e um sumário é obtido como saída (NENKOVA & MCKEOWN, 2011).

À vista de que a SAT é um processo, um sumário pode ser definido, em termos gerais, como sendo um texto produzido por meio de **um ou mais textos**, que transmite **informações importantes** do(s) texto(s) de origem e que normalmente **é menor que a metade** do(s) texto(s) original(s) (RADEV; HOVY; MCKEOWN, 2002).

Os elementos grifados na definição de sumário caracterizam os principais aspectos na pesquisa em SAT:

- Sumários podem ser produzidos de um único documento ou conjunto de documentos;
- Sumários devem preservar informações importantes;
- Sumários devem ser curtos.

A definição de sumário apresentada é genérica e de granularidade pouco refinada.

Ao longo do tempo algumas taxonomias ou formas de classificação para sumários foram apresentadas pela comunidade científica, tais como as propostas pelos autores: Spärck Jones, Hovy e Lin e Mani e Maybury. Assim como ocorre em várias áreas de pesquisa, a ST está em constante evolução, e novas abordagens vêm surgindo, o

que torna difícil classificar um sistema de sumarização, pois novas características são estudadas e publicadas.

O Quadro 1 apresenta um resumo dos tipos de sumários e quais são os fatores expressivos para se levar em consideração na classificação.

Quadro 1 – Classificação do processo de sumarização conforme alguns fatores

FATOR	TIPO
Mídia	Texto
	Imagem
	Vídeo
	Fala
	Hipertexto
Entrada	Um documento (*single-document* ou monodocumento)
	Conjunto de documentos (*multi-document*s ou multidocumentos)
Saída	Extrato
	Abstrato
	Headline
Finalidade	Genérico
	Crítico
	Indicativo
	Informativo
	Personalizado
	Focado em consulta
	Atualização
	Baseado no sentimento
Linguagem	*Mono-lingual*
	Multi-lingual
	Cross-lingual

Fonte: Adaptado de Lloret e Palomar (2012)

Tradicionalmente, o processo de sumarização tem se voltado à informação na forma de texto, contudo podem-se utilizar informações multimídias, por exemplo: imagem, vídeo, áudio, textos *online* e hipertextos.

Com relação à entrada, um sumário pode ser originado de um ou vários documentos e, em relação a saída, pode ser um extrato – quando as frases significativas são selecionadas e apresentadas, seguindo uma ordenação; abstrato – quando as frases são selecionadas e novos vocabulários são adicionados, podendo ser um substituto do texto

original e *headline* – são frases que servem de título ou subtítulo (LLORET; PALOMAR, 2012).

Um sumário é genérico quando tenta representar todos os fatos relevantes, podendo ser um substituto do texto-fonte. Sumário crítico ou avaliativo tem por objetivo expressar pontos de vista do autor sobre um assunto específico, que inclui comentários, sugestões, recomendações. Sumários indicativos são utilizados para assinalar quais são os tópicos abordados, dando uma noção do que é o texto-fonte, como os índices dos livros, por exemplo, mas não substituem o texto original (CAMARGO, 2013). Os sumários informativos apresentam os principais tópicos, dispensando a leitura do texto-fonte, pois fornecem informações mais detalhadas (LLORET; PALOMAR, 2012). O objetivo do sumário personalizado é fornecer um resumo contendo informações específicas com base no perfil do usuário (MÓRO; BIELIKOV, 2012). Sumários com foco em consulta apresentam um conteúdo que é conduzido pela necessidade de um usuário ou consulta (LUO et al., 2013).

Os sumários de atualização e sentimento estão relacionados ao surgimento da Web 2.0 que tem incentivado novos tipos de gêneros textuais.

Principalmente por meio das redes sociais, qualquer pessoa pode expressar seus sentimentos com relação a um tema/entidade, produto ou serviço, o que permitiu surgir os sumários baseados no sentimento. O sumário de atualização assume que o usuário já tem algum conhecimento sobre determinado tema, portanto apenas as informações mais recentes são mostradas (LLORET; PALOMAR, 2012).

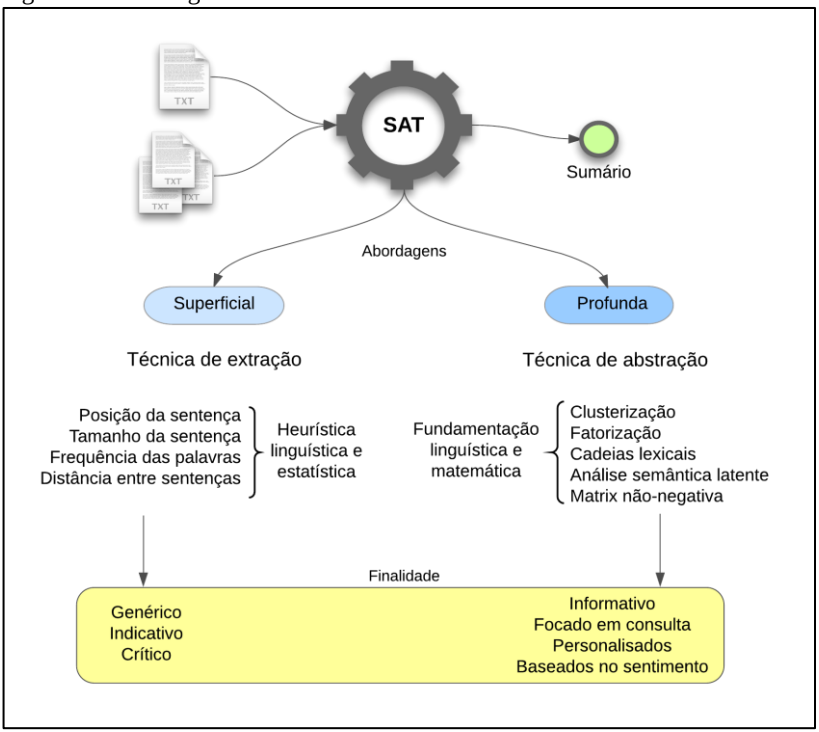
Finalmente, dependendo do número de línguas na modalidade escrita em questão, um sumário pode ser *mono-lingual* – quando a linguagem de entrada e saída for a mesma e o sistema trabalhar em uma única língua, *multi-lingual* ou *cross-lingual* – quando diferentes línguas estão envolvidas.

Por exemplo, um sistema de sumarização que somente produz sumários em Português limita os textos de entrada a serem em língua portuguesa, portanto é *mono-lingual*. Se o sistema aceitar mais de uma língua, por exemplo, Português, Espanhol ou Inglês, mas produz sumários na mesma língua de entrada, é o caso de *multi-lingual*. Se o resumo está em Português, mas o texto-fonte está em Inglês, a sumarização é *cross-lingual*, pois as línguas de entrada e saída são diferentes.

2.1.1 Abordagens da SAT

Segundo Jones (2007) e Castro Jorge e Pardo (2010), existem duas abordagens clássicas para o problema de sumarização dentro da área de PLN: a superficial com o predomínio da técnica de extração e a profunda onde se encontra a técnica de abstração, representadas na Figura 2.

Figura 2 – Abordagens e técnicas da SAT



Fonte: Elaborada pelo autor

A abordagem superficial utiliza métodos estatísticos e/ou empíricos para obter o sumário. Essa abordagem se caracteriza pela extração de sentenças copiadas do texto-fonte, por isso é denominada como técnica de extração (BALAGE FILHO; PARDO; NUNES, 2007) (BATCHA; ZAKI, 2010). As sentenças são extraídas com base na frequência das palavras, posicionamento e comprimento em relação ao texto.

A abordagem profunda utiliza técnicas formais e modelos linguísticos. Para construir um sumário segundo essa abordagem, o texto-fonte é analisado semanticamente com base na identificação da relação entre as sentenças e entre as palavras. O sumário gerado é uma versão reformulada do texto original, e não apenas uma cópia. Essa técnica é denominada de abstrativa (DAS; MARTINS, 2007; BATCHA; ZAKI, 2010).

Palavras que não estão no vocabulário do texto-fonte podem ser incluídas quando o sumário é abstrativo, ou ainda fusão de sentenças redundantes quando a sumarização é de múltiplos documentos. A fusão de sentenças envolve a identificação e combinação de frases que transmitem informações semelhantes, em uma frase (BARZILAY; MCKEOWN, 2005; SUNEETHA, 2011).

Fontes de informação da heurística linguística e estatística, tais como: posição e comprimento da sentença, frequência das palavras, distância entre sentenças são usados predominantemente em métodos não-supervisionados[2].

Abordagens estatísticas são eficientes em termos de computação, enquanto que abordagens linguísticas por tratarem da semântica dos termos, podem produzir melhores sumários (CHANDRA; GUPTA; PAUL, 2011).

A Figura 1 apresenta as principais técnicas da SAT com os seus respectivos métodos, mas não existe exclusividade em relação aos métodos. Embora métodos estatísticos sejam empregados para gerar sumários pela técnica extrativa, nada impede que sejam utilizados em conjunto com métodos de fundamentos linguísticos para gerar sumários pela técnica abstrativa.

2.1.2 Técnica de extração

A sumarização extrativa é formada pela seleção, extração e reorganização dos segmentos mais informativos de um texto, onde o primeiro passo é identificar quais são as características textuais que podem ser consideradas para definir a importância das sentenças. O sumário extrativo é mais simples de ser produzido e o processo de

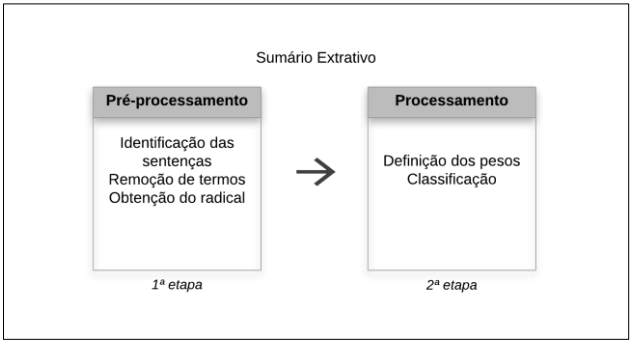
[2] Métodos não-supervisionados de extração de sentença são os que dispensam anotação humana, não dependem de nenhum conhecimento fonte, modelo externo ou processamento e interpretação linguística (NENKOVA; MCKEOWN, 2011).

sumarização pode ser dividido em duas etapas: pré-processamento e processamento, conforme a Figura 3.

A etapa de pré-processamento, geralmente inclui: identificação das sentenças, eliminação de termos que não agregam informações relevantes e obtenção do radical de cada palavra.

Na etapa de processamento, as características que influenciam na relevância das sentenças são calculadas e atribui-se um peso. As sentenças são ranqueadas sob uma determinada ordem e classificadas.

Figura 3 – Etapas da sumarização extrativa



Fonte: Elaborada pelo autor

A seguir são descritas as principais abordagens de sumarização extrativa.

2.1.2.1 Frequência das palavras

O primeiro trabalho na área que explorou o uso da frequência lexical como um indicador de importância foi conduzido por Luhn, na década de 50. Luhn apresentou a ideia de que algumas palavras, em um documento, são descritivas do seu conteúdo, e as frases mais importantes são as que contêm muitas dessas palavras próximas umas das outras (LUHN, 1958).

Ao elaborar um texto, um escritor usualmente repete certas palavras conforme desenvolve o seu raciocínio. Luhn sugeriu que a frequência de ocorrência das palavras descritivas (também chamadas de palavras significativas, palavras conteúdo, palavras-chaves ou

keywords)[3] em um texto pode ser útil no cálculo da relevância das sentenças (fator de significância).

Contudo, há duas ressalvas feitas por Luhn em sua abordagem: algumas das palavras mais frequentes em um texto não são palavras descritivas, assim como as palavras que aparecem poucas vezes (NENKOVA; MCKEOWN, 2011).

Palavra cuja classe gramatical é pronome, preposição, artigo é desprovida de valor semântico, portanto não serve para indicar o tema de um documento. O mesmo acontece com palavras de um domínio específico, por exemplo, a palavra "variável" em um texto científico da área da computação.

Luhn propôs o uso de uma lista pré-definida, chamada de *stop word list*, constituída por palavras de significado irrelevante a fim de serem removidas do texto. Luhn também usou limiares de frequência para filtrar as palavras que ocorrem pouco e as que ocorrem muito. As palavras que estiverem dentro dos limites mínimo e máximo de frequência são consideradas palavras-chave (LUHN, 1958).

O critério para produzir a pontuação de cada sentença leva em consideração a relação entre as palavras descritivas e não a distribuição. Para cada sentença, grupos (*clusters*) de 4 a 5 palavras não significantes delimitados por duas palavras descritivas são formados.

O fator de significância da sentença é determinado pelo quadrado do número de palavras significantes contidas no *cluster*, dividido pelo número total de palavras do *cluster*. Por exemplo, se uma sentença apresenta um *cluster* de 7 palavras sendo que 4 são palavras descritivas, 2,3 é o seu fator de significância.

Se houver sobreposição de *clusters*, aquele com o maior fator de significância será preferido à pontuação da frase. As sentenças com maior pontuação são as que devem compreender o sumário.

A frequência lexical como um indicador de importância pode ser usada para calcular a **probabilidade**, ***tf-idf***, **função de verossimilhança** (NENKOVA; MCKEOWN, 2011).

a) Probabilidade das palavras

Dado um documento ou conjunto de documentos, a probabilidade P é calculada a partir do número de ocorrências de uma palavra np dividido pelo número de todas as palavras da entrada N.

[3] São palavras tais como: verbos, substantivos, adjetivos e números.

$$P(p) = {}^{np}/_{N} \quad (1)$$

A pontuação de uma sentença considera a probabilidade média das palavras descritivas que a constitui.

A probabilidade de uma palavra é a forma mais simples de pontuar uma sentença, porém eficiente. O SumBasic é um sistema de sumarização multidocumento, baseado na abordagem de algoritmos gulosos, onde a probabilidade de palavras foi implementada primeiramente (YIH et al., 2007).

O sistema foi testado usando dois conjuntos de documentos. Um pertencendo à conferência DUC 2004 e outro a MSE 2005. Em ambos os casos foi utilizada a medida ROUGE-1, e o SumBasic ficou entre os melhores sistemas avaliados (YIH et al., 2007; NENKOVA; VANDERWENDE, 2005).

Segundo Nenkova e Vanderwende (2005), a alta frequência das palavras é um dos fatores que afeta a decisão de um humano para incluir conteúdo específico em um sumário. Esse fato foi comprovado pelas autoras quando analisaram os sumários gerados por humanos e máquinas do conjunto de dados da DUC 2003. Os sumários humanos apresentaram maiores valores de probabilidades de palavras descritivas em relação aos gerados por máquinas.

b) Tf-idf

Um dos problemas em se utilizar a frequência das palavras como um indicador de importância é o fato de algumas palavras aparecerem muitas vezes no texto, enquanto outras palavras raramente aparecem.

Determinar um ponto de corte na frequência das palavras não é algo trivial. Sendo assim, pesquisadores optaram por elaborar uma lista com as palavras mais frequentes de uma linguagem e palavras de domínio, que não acrescentam importância na ponderação de uma sentença. As palavras mais frequentes são conhecidas como *stop words*.

Na área de Recuperação da Informação (RI), a medida *tf-idf* (*Term Frequency-Inverse Document Frequency*) é bastante fundamentada e empregada para contar palavras de grandes coleções de documentos (NENKOVA; MCKEOWN, 2011). O peso *tf-idf* pode ser calculado da seguinte forma:

(2)

$$tf\_idf_w = tf \times idf = tf \times \log \frac{D}{d(p)}$$

A frequência de uma palavra *tf* é definida como sendo o número de vezes que a palavra aparece no documento *d*. Entretanto, a frequência deve ser normalizada para prevenir um viés em documentos longos, a normalização *idf*.

A frequência inversa da palavra *idf* é uma medida geral de importância da palavra, sendo definida pelo logaritmo do quociente entre o número total de documentos *D* e o número total de documentos *d* que contém a palavra *p*.

A normalização da frequência diminui o peso das palavras que ocorrem com mais frequência na coleção. Isso reduz a importância de palavras comuns à coleção, garantindo que os correspondentes documentos sejam mais influenciados por palavras descritivas contendo uma frequência relativamente baixa na coleção (AGGARWAL; ZHAI, 2012).

A normalização da frequência de palavras em documentos é em si uma vasta área de pesquisa, e uma variedade de outras técnicas que discutem diferentes métodos podem ser encontrados na literatura (LI; YAMANISHI, 1997).

Se a noção de documento na medida *tf-idf* for substituída pela de sentença, tem-se a medida *tf-isf* (*Term Frequency-Inverse Sentence Frequency*). A sentença é pontuada conforme a média aritmética do peso *tf-isf* de todas as suas palavras.

A medida *tf-idf*, seja de uma forma ou de outra, é uma das mais empregadas em sumarização extrativa, principalmente por ser rápida e fácil de calcular.

Doko, Štula e Šerić (2013) relatam sobre esse fato quando, em seu trabalho, propõem variantes da medida *tf-isf* com base em critérios locais de vizinhança e comprimento das sentenças. Os melhores resultados são obtidos quando combinados os critérios locais.

c) Função de verossimilhança

Segundo Nenkova e McKeown (2011), a razão de verossimilhança é um teste estatístico utilizado para comparar o ajuste dos dados sob dois modelos, o nulo e o alternativo . O teste baseia-se na relação de probabilidade que expressa o quão provável os dados estão sob um dos modelos.

Em muitas aplicações é conveniente trabalhar-se com o log natural da função de verossimilhança, chamado de log-verossimilhança (*log-likelihood).* O log natural da função de verossimilhança pode ser usado para calcular um valor de *p* que decide sobre rejeitar o modelo nulo em favor do modelo alternativo, ou não.

O log-verossimilhança é uma forma de identificar palavras descritivas. Enquanto *tf-idf* atribui peso a todas as palavras de entrada, o log-verossimilhança divide as palavras em duas classes – as que são descritivas e as que não são (NENKOVA; MCKEOWN, 2011).

Para estimar as classes é necessário um grande corpus de fundo. O corpus é responsável por estabelecer a relação mútua entre uma palavra e uma entrada a ser sumarizada (GUPTA; NENKOVA; JURAFSKY, 2007).

A verossimilhança entre uma entrada θ e o corpus D é definida nas seguintes hipóteses de probabilidade:

(3)

$$H_1 : P(w|\theta) = P(w|D) = p;\ isto\ \text{é}\ (w\ não\ é\ uma\ palavra\ descritiva)$$
$$H_2 : P(w|\theta) = p_\theta\ e\ \ P(w|D) = p_D;\ \ p_\theta > \ p_D\ ; isto\ \text{é}\ (w\ \text{é}\ descritiva)$$

A hipótese H_1 sugere que a probabilidade de uma palavra w na entrada θ seja a mesma que no corpus D. A hipótese H_2 assume que a probabilidade de uma palavra é diferente, maior na entrada do que no corpus (LIN; HOVY, 2000).

A probabilidade total de que a palavra w apareça k vezes nos N ensaios segue o modelo de distribuição binomial definido por:

(4)

$$f(k; N, p) = \binom{N}{p} p^k (1-p)^{N-k}$$

Conforme as hipóteses H_1 e H_2, a razão de verossimilhança λ pode ser definida como sendo:

(5)

$$\lambda = \frac{verossimilhança\ das\ informações\ dado\ H_1}{verossimilhança\ das\ informações\ dado\ H_2}$$

Uma propriedade útil da razão do log-verossimilhança é que a medida $-2\,log\,\lambda$ tem distribuição muito próxima da distribuição qui-quadrado (χ^2). Isso significa que uma palavra aparece significativamente mais frequente na entrada do que no corpus quando $-2\,log\,\lambda > 10$, pois um valor de 10,83 pode ser obtido na distribuição χ^2 com *N – 1* graus de liberdade (LIN; HOVY, 2000).

Uma sentença é pontuada conforme o número de palavras descritivas que apresenta. As palavras descritivas também são chamadas de palavras tópico (*topic words)*, assinaturas tópica (*topic signatures)* ou termos assinatura (*signature terms*[4]*)*.

Lin e Hovy (2000) foram os primeiros a utilizar a ponderação de probabilidades à ST. A cada palavra descritiva, denominadas de *signature terms*, foi atribuído o peso igual a 1, enquanto que o restante das palavras receberam peso igual a 0. O peso de uma sentença de entrada foi definido como sendo o somatório de suas palavras descritivas.

Em Gupta, Nenkova e Jurafsky (2007) o método de probabilidade das palavras é comparado com o log-verossimilhança (em três variações). A ponderação das sentenças seguiu a proposta por Lin e Hovy. Os resultados demostraram que para sumários genéricos, não existiu diferença significante entre os métodos, porém, quando o sumário é focado em consulta, duas variações do log-verossimilhança levaram vantagem. Em Katragadda (2010) o log-verossimilhança é utilizado na avaliação automática de sumários.

2.1.2.2 Posição da sentença

Ainda na década de 50, Baxendale publicou um trabalho sobre um recurso simples e útil para encontrar pontos relevantes no texto: a posição da sentença.

O autor analisou 200 parágrafos de textos científicos e descobriu que, em 85% dos casos, a sentença mais importante do parágrafo era a primeira e em 7%, a última. Assim, uma forma precisa de selecionar a principal sentença seria escolhendo uma das duas opções. Desde então este recurso vem sendo usado em muitos sistemas baseados em aprendizagem de máquina complexos (DAS; MARTINS, 2007).

2.1.2.3 Frequência das palavras e posição das sentenças

Outro trabalho que também fundamentou várias outras pesquisas em ST foi o de Edmundson, na década de 60. Edmundson ampliou a abordagem de Luhn propondo que várias podem ser as características que definem a importância de uma sentença.

[4] É como são chamadas as palavras descritivas em alguns trabalhos da área e que fazem uso de função de verossimilhança.

A principal contribuição do trabalho de Edmundson foi o desenvolvimento de uma estrutura típica para a sumarização extrativa (BALABANTARAY et al., 2012). O sistema produz extratos indicativos e foi baseado em quatro métodos básicos chamados de Sinal (*Cue*), Chave (*Key*), Título (*Title*) e Localização (*Location*) (EDMUNDSON, 1969).

a) Sinal: considera a hipótese de que a relevância de uma sentença é influenciada pela presença de certas palavras pragmáticas - substantivos superlativos, advérbios de conclusão e termos de causalidade tais como: "*significant*", "*impossible*" e "*hardly*". Um dicionário de palavras pré-armazenadas, compiladas com base em dados estatísticos de um corpus, é utilizado para identificar as palavras sinais. O dicionário de sinais é dividido em três subdicionários: palavras positivamente relevantes, palavras negativamente relevantes e palavras irrelevantes. O peso Sinal de cada frase é a soma dos pesos de suas palavras sinais constituintes.

b) Chave: o princípio do método Chave é o mesmo proposto por Luhn, de que a alta frequência de certas palavras pode demonstrar o conteúdo do texto. Para cada documento é compilado um glossário chave, consistindo de palavras estatisticamente selecionadas do próprio documento. As palavras são ordenadas da maior frequência acumulada para a menor, e as que encontrarem-se acima de um determinado limiar de frequência são consideradas palavras-chave. A palavra-chave recebe peso positivo igual a sua frequência no documento. O peso Chave da sentença é a soma dos pesos de suas palavras-chave constituintes.

c) Título: o método de Título considera a hipótese de que quando um autor particiona o corpo do documento em seções, ele o faz por meio de títulos e subtítulos compostos de palavras que descrevem o conteúdo da seção. Para cada documento é compilado um glossário título, contendo todas as palavras relevantes do título, subtítulos ou cabeçalhos. As palavras do título podem receber um peso maior do que as palavras de subtítulos e cabeçalhos. O peso Título de uma sentença é a soma dos pesos das palavras do título que estão presentes na sentença.

d) Localização: o método Localização é baseado em duas hipóteses. A primeira refere-se ao título de seções enquanto que a segunda, ao formato do documento. As sentenças que apresentam certas palavras de títulos de seções são positivamente relevantes, assim como as sentenças que tendem a ocorrer ao início ou ao final do documento ou do parágrafo. O método Localização usa um dicionário de palavras pré-armazenado, de um corpus, que aparecem em títulos de seções, por exemplo: "*Introduction*", "*Purpose*" e "*Conclusions*". O método atribui

um peso positivo para a sentença que apresentar tais palavras, assim como atribui um aumento de peso à sentença que esteja no primeiro ou no último parágrafo, ou ainda, seja a primeira ou a última sentença de qualquer outro parágrafo.

Ao final, as sentenças são pontuadas conforme a combinação linear de funções de peso, dos quatro referidos métodos: Sinal (S), Chave (C), Título (T) e Localização (L).

$$p = a_1 S + a_2 C + a_3 T + a_4 L \quad (6)$$

Os parâmetros a_1, a_2, a_3 e a_4 são ajustados manualmente. Edmundson sugere em seus resultados que a frequência das palavras-chave (método Chave) é o menos importante dos métodos citados. A combinação dos outros três métodos foi a que melhor aproveitou o conhecimento do domínio para sumarização (EDMUNDSON, 1969).

Kupiec, Pedersen e Chen (1995) basearam-se na pesquisa de Edmundson para propor que o ajuste manual dos pesos da equação 6 fosse transformado em um problema de aprendizado de máquina, utilizando um classificador Naïve Bayes.

Embora o trabalho de Edmundson tenha influenciado várias pesquisas subsequentes em ST, seu método apresenta a desvantagem de ter usado apenas características superficiais simples e não ter levado em consideração uma taxa de compressão na geração dos sumários (ANTIQUEIRA, 2007).

Outro fato a se mencionar é com relação ao corpus de textos científicos utilizados no estudo. Os resultados podem não ser válidos se outro corpus de textos for utilizado, pois pode-se configurar uma nova importância nos métodos empregados para pontuar o peso das sentenças (ANTIQUEIRA, 2007).

2.1.2.4 Frases indicativas

Certas passagens do texto formam estruturas que são indicadores importantes do seu assunto. Paice (1980) chamou essas estruturas de frases indicativas (*indicators)*. Frases como, "*The principal aim of this paper is to investigate...*" e "*In the present paper, a method is described for...*", são exemplos de frases indicativas em textos científicos.

Paice observa que há casos em que um documento pode conter nenhuma a várias frases indicativas, portanto a abordagem de frases indicativas deve fazer uso de outros atributos para a seleção de sentenças.

Um modelo, semelhante a uma árvore, agrega as palavras pertencentes ao mesmo grupo e que formam as frases indicativas. O modelo é utilizado para a atribuição de pesos e, assim selecionar os trechos com melhor pontuação para compor um sumário indicativo.

O principal problema da abordagem de frases indicativas é o modelo utilizado na pontuação. Para representar todos os possíveis indicadores é necessário milhares de palavras, o que pode deixar o modelo grande e o processo lento (PAICE, 1980). Além disso, torna o sistema dependente da linguagem dos textos a serem sumarizados, dificultando a portabilidade.

2.1.3 Avaliação de sumários

Produzir um sumário automaticamente é algo desafiador. Segundo Lloret e Palomar (2012), questões como a redundância, dimensão temporal, correferência ou ordenação de sentença, para citar alguns, devem ser levadas em consideração.

Para Mani (2001) e Jones (2007), genericamente, os métodos de avaliação de sumarização de texto são classificados em intrínsecos e extrínsecos. O primeiro avalia a coerência e informatividade de sumários – testa o sistema em si; enquanto que o segundo avalia o impacto da sumarização com relação a tarefas como: a compreensão da leitura, avaliação da relevância, entre outros.

Avaliações intrínsecas são normalmente utilizadas em ambos os casos: quando é necessário o julgamento humano sobre determinado sumário, ou em casos que requerem uma comparação entre o sumário originado automaticamente com outro sumário de autoria humana – padrão-ouro (*gold standard*) (NENKOVA; MCKEOWN, 2011). Avaliações extrínsecas também tentam medir a qualidade do sumário, mas por meio de tarefas orientadas a recuperação de informação.

2.1.3.1 Abordagens de avaliações intrínsecas

As avaliações intrínsecas tomam a forma manual, semi-automatizada ou automatizada. Na avaliação manual uma ou mais pessoas avaliam o sumário produzido por um sistema. Na avaliação automática o sumário é avaliado por um sistema, sem a interferência humana, e na semi-automática a avaliação dá-se por um sistema, mas os resultados dependem da interpretação humana e do domínio do conhecimento. Em ambos os casos há o julgamento humano e uma variabilidade decorrente (KATRAGADDA, 2010).

Algumas das principais abordagens para avaliação intrínseca de sumários são descritas a seguir.

a) Precisão, Cobertura e Medida-f (*Precision, Recall and F-measure*): no caso da sumarização extrativa, uma pessoa seleciona as frases mais importantes do texto-fonte para compor o sumário padrão. As sentenças do sumário produzido pelo sistema sumarizador são comparadas com as sentenças do sumário padrão por meio das métricas de RI – precisão, cobertura[5] e medida-f, definidas como sendo:

$$precisão = \frac{quant.de\ sentenças\ sobrepostas\ entre\ os\ sumários}{quant.de\ sentenças\ selecionadas\ pelo\ sistema} \quad (7)$$

$$cobertura = \frac{quant.de\ sentenças\ sobrepostas\ entre\ os\ sumários}{quant.de\ sentenças\ selecionadas\ pela\ pessoa} \quad (8)$$

$$medida\text{-}f = \frac{2 \times precisão \times cobertura}{precisão + cobertura} \quad (9)$$

A medida-f é a combinação da precisão e cobertura em uma única métrica que representa o desempenho de um sistema.

Nenkova e McKeown (2011) afirmam que as medidas de precisão e cobertura apresentam alguns problemas, destacando: o julgamento humano, a granularidade e a semântica.

Diferentes pessoas tendem a escolher diferentes sentenças na construção do sumário padrão. Portanto, um sistema pode selecionar uma boa sentença, mas ainda assim ser penalizado na avaliação das medidas porque tal sentença não consta no sumário padrão. Analisando também o fato de que a precisão é uma medida mais rigorosa, a cobertura pode vir a ser mais importante na avaliação de sumários.

A granularidade definida pela sentença também representa um problema. Sentenças diferem em quantidade de palavras, sendo que em alguns momentos pode ser mais informativo a escolha de uma sentença longa em vez da curta.

Com relação à semântica, duas frases distintas podem expressar o mesmo significado, algo comum na sumarização de conjunto de documentos. Como apenas uma das sentenças está no sumário padrão, o sistema será penalizado caso selecione uma sentença equivalente.

[5] Ambas são utilizadas para estimar o desempenho de um sistema de RI, porém a precisão é uma medida de exatidão e a cobertura é de completude.

Para contornar esses problemas, alguns autores sugerem utilizar mais de um sumário padrão, maior atenção à medida de cobertura e menores unidades de análise orientadas semanticamente.

b) Utilidade relativa (*Relative Utility*): juízes são convidados a atribuir uma pontuação numérica para as frases do documento de entrada, em uma escala de 0 a 10. A pontuação 10 indica uma sentença central para o tema do documento, altamente adequado sua inclusão no sumário, enquanto a pontuação 0 marca uma frase como irrelevante (RADEV; TAM; ERKAN, 2003).

A utilidade relativa foi validada no trabalho de Radev, Tam e Erkan (2003) como sendo uma métrica de avaliação de sumários, capaz de tratar os problemas de variabilidade do julgamento humano e equivalência semântica, existentes na avaliação com precisão e cobertura.

Contudo, Nenkova, Passonneau e McKeown (2007) consideram que embora a abordagem seja atraente e intuitiva, requer esforço manual para a pontuação das frases. Além disso, a utilidade relativa pode não distinguir entre sumários humanos e automáticos, pois os sumários automáticos podem atingir uma pontuação maior do que os produzidos por humanos.

c) ROUGE[6] (*Recall-Oriented Understudy for Gisting Evaluation*): é uma ferramenta para avaliar a semelhança entre sumários produzidos automaticamente e manualmente (LIN, 2004). A técnica implementada em ROUGE é derivada do método BLEU (*Bilingual Evaluation Understudy*) usada na avalição automática de sistema de tradução de máquina (LLORET; PALOMAR, 2012).

ROUGE analisa a sobreposição de elementos básicos (tais como n-gramas, sequências ou pares de palavras) do sumário automático e do modelo. Dessa forma, compara-se o conteúdo do sumário automático com o de um ou mais sumários padrões, verificando a co-ocorrência de palavras comuns a todos (KATRAGADDA, 2010).

O pacote de avaliação ROUGE obtém a precisão, cobertura e medida-f e, quatro diferentes tipos de medidas de n-gramas: ROUGE-N (unigrama quando N = 1, bigrama quando N = 2) compara os n-gramas de dois sumários, e conta o número de sobreposições; ROUGE-L para a comparação de sequência de palavras longas, ROUGE-W para a ponderação da subsequência comum mais longa e ROUGE-S para a comparação de bigramas em sequências arbitrárias (ALGULIEV et al., 2011).

[6] http://www.berouge.com/Pages/default.aspx

Segundo Lin (2004), ROUGE-N é uma medida de cobertura definida pela fórmula:

$$ROUGE - (N) = \frac{\sum_{S \in \{Refs\}} \quad \sum_{n-grama \in S} total_{sobreposição}\,(n-grama)}{\sum_{S \in \{Refs\}} \quad \sum_{n-grama \in S} total\,(n-grama)} \tag{10}$$

Onde, N é o tamanho de n-grama, S é um sumário, $Refs$ é o conjunto dos sumários de referência (sumário padrão), $total_{sobreposição}(n-grama)$ é o número de n-gramas que co-ocorrem no sumário S e no conjunto de sumários de referência e $total\,(n-grama)$ é o número de n-gramas do sumário de referência.

ROUGE-N é uma métrica amplamente aplicada na avaliação de sumários. Nenkova e McKeown (2011) afirmam que ROUGE é comumente utilizada para avaliar sumários porque é de baixo custo e rápida. Avaliações manuais são demoradas, caras e de difícil repetição. A automatização de métricas de avaliação de sumários é importante para manter o controle de melhorias em sistemas.

d) Método Pirâmide (*Pyramid Method*): foi proposto por Nenkova e Passonneau (2004) abordando o problema da variação da seleção de conteúdo em sumários humanos e da dependência dos resultados em relação ao modelo usado para a avalição.

O método baseia-se em identificar as informações de mesmo significado em diferentes sumários de autoria humana, chamadas de unidades de conteúdo semântico (*Semantic Content Units* - UCS), a fim de obter o padrão-ouro para a avaliação (NENKOVA; PASSONNEAU; MCKEOWN, 2007; PASSONNEAU, 2009).

Cada UCS recebe um peso igual ao número de avaliadores humanos que expressaram o UCS em seus sumários. Os pesos seguem uma distribuição específica que permitem diferenciar o conteúdo importante do menos importante.

O novo sumário padrão com *n* unidades de conteúdo é definido pela razão entre a soma dos pesos do conteúdo expresso em um sumário e a soma dos pesos do sumário ideal com o mesmo número de UCSs. O escore obtido varia entre 0 e 1.

Um inconveniente no método é a necessidade de um grande esforço manual para anotar as UCS ao longo de uma coleção de sumários (LLORET; PALOMAR, 2012). Além disso, o método foi desenvolvido para a avaliação de sumários abstrativos, requerendo uma

análise desnecessária para o caso de sumários extrativos (NENKOVA; MCKEOWN, 2011).

Pesquisas publicadas em conferências da área têm sugerido novos modelos para avaliar a informatividade de sumários, prevalecendo a forma automatizada. Amigó et al. (2005) propõem um framework de avaliação (QARLA) com medidas provenientes da análise da similaridade entre um conjunto de sumários de referência e sumários automáticos.

O BEwT-E[7] (*Basic Elements with Transformations for Evaluation*) é a implementação do framework BE (*Basic Elements*) para avaliar sumários. O intuito do BE é decompor um sumário em pequenas unidades de conteúdo, chamadas de elementos básicos e, em seguida, compará-las para obter um escore de similaridade (TRATZ; HOVY, 2008).

Giannakopoulos et al. (2008) sugerem o AutoSummENG (*AUTOmatic SUMMary Evaluation based on N-gram Graphs*), uma ferramenta que compara a representação gráfica de n-gramas dos resumos automáticos com os resumos de referência. Devido a sua natureza estatística, a abordagem apresentada é uma linguagem neutra, o que permite trabalhar em outras línguas.

Outro método automático é o GEMS (*Generative Modeling for Evaluation of Summaries*), proposto por Katragadda (2010), que sugere o uso de termos de assinatura. Os termos de assinatura são identificados nos resumos de referência com base em um modelo de consistência e *tags parts-of-speech* (POS), tais como verbos ou substantivos. A distribuição dos termos de assinatura do texto-fonte é calculada e utilizada para obter a probabilidade de um sumário estar sendo influenciado por tais termos.

2.1.3.2 Avaliação da qualidade

As abordagens e sistemas apresentados no item anterior avaliam a qualidade de sumários conforme o seu conteúdo. Há também abordagens que se preocupam em avaliar a qualidade de sumários considerando outros aspectos importantes da informatividade.

Nas conferências DUC e TAC os sumários são avaliados considerando questões linguísticas de qualidade: gramaticabilidade, redundância, clareza referencial, foco, estrutura e coerência.

[7] http://www.isi.edu/publications/licensed-sw/BE/index.html

Avaliações desse tipo dispensam a necessidade de um sumário padrão e são feitas manualmente, por especialistas humanos, que marcam a qualidade do sumário seguindo uma escala de 1 a 5, para cada uma das questões.

Um exemplo de avaliação da qualidade de sumários é proposta em Vadlapudi e Katragadda (2010). Os autores abordam o problema da identificação do grau de aceitação das formações gramaticais no nível de sentença, utilizando características superficiais, como n-gramas e sequências LCS (*Longest Common Subsequence*) para avaliar a estrutura e coerência.

2.1.4 Considerações sobre SAT

Embora a tentativa de gerar textos automaticamente tenha iniciado há 50 anos, foi apenas nos últimos anos que a SAT experimentou um grande desenvolvimento[8].

Várias publicações propondo métodos sofisticados e abordagens combinadas têm surgido para tratar desse campo de pesquisa. Quando analisada a literatura sobre ST, fica evidente o predomínio da técnica extrativa sobre a abstrativa. Contudo, esse domínio maior de uma técnica tende a diminuir a medida que novas pesquisas e tecnologias são apresentadas pela comunidade científica.

Nos últimos anos, têm sido propostos vários métodos de sumarização multidocumento, baseados em grafos e aprendizado de máquina (BARZILAY; MCKEOWN, 2005) (YIH et al., 2007) (ALGULIEV; ALYGULIEV, 2008) (GIANNAKOPOULOS et al., 2008) (MEI; CHEN, 2011) (ALGULIEV et al., 2011) (RIBALDO; AKABANE, 2012) (LUO et al., 2013) (ARIES; OUFAIDA; NOUALI, 2013).

Métodos baseados em grafos consistem, basicamente, em representar de forma gráfica uma topologia capaz de revelar como os elementos do texto se relacionam. Normalmente, os nós são palavras, frases ou parágrafos, enquanto que as arestas são as ligações entre os elementos (concebidas por relações semânticas ou medidas de similaridade, dependendo do propósito).

[8] Um dos fatores responsáveis pelo desenvolvimento das pesquisas em ST foi o surgimento da Internet e, com ela, a grande quantidade de material textual publicado.

Abordagens baseadas em aprendizado de máquina baseiam-se em algoritmos Bayesianos, Modelos de Cadeias de Markov, Redes Neurais, Modelos de Regressão e precisam de um corpus de treinamento.

Das e Martins (2007) afirmavam que o tema Avalição era crucial e que certamente conduziria muitas pesquisas. Segundo Lloret e Palomar (2012), a avaliação de sumários é um desafio porque não está claro, mesmo por seres humanos, que tipo de informação um sumário deve conter. Dependendo da aplicação a que se destina, a informação nem sempre será a mesma.

A avaliação automatizada de sumários é mais complexa do que inicialmente parecia ser. As formas de avaliações, principalmente por meio de padrão-ouro, medidas de extração de informações, e que foram úteis para tarefas como a tradução automática, são menos indicadas para expressar o potencial de sumários. Ao mesmo tempo, é difícil uma forma de avaliação discriminar sobre as particularidades entre os sistemas quando se compara sumários de máquinas com sumários humanos.

Apesar da complexidade do processo de avalição, as formas automatizadas que comparam o conteúdo de sumários, tais como as abordagens correspondentes a n-gramas, são as mais empregadas, pois demonstraram alta correlação com as avaliações manuais.

Portanto, sobre avaliação é plausível afirmar – sistemas de avaliação de sumários que contemplem as especificidades da área serão possíveis se existirem sistemas sumarizadores capazes de lidar com a variabilidade do julgamento humano na composição de sumários.

Um dos fatores responsáveis por avanços significativos na tecnologia de sumarização foram as conferências DUC/TAC. A DUC (*Document Understanding Conferences*) era uma conferência destinada somente a ST e foi realizada anualmente de 2001 a 2007. Desde 2008, a DUC passou a fazer parte da TAC (*Text Analysis Conference*), a qual incluiu uma trilha específica para ST e se destina a áreas com foco em PLN.

Para se adaptar aos novos desafios e exigências, as tarefas envolvidas nas conferências foram mudando ao longo das edições. No início, a sumarização era mono e multidocumento, de notícias e destinada a produzir sumários genéricos. Nas últimas edições predominou a sumarização multidocumento, com a inclusão de novos tópicos: sumarização *cross-lingual*, focada em consulta, de atualização, baseada no sentimento (*blogs*), avaliação.

A TAC, edição 2014, trouxe a trilha de Sumarização Biomédica (*BiomedSumm*). A ideia da tarefa na trilha é a de que um conjunto de

sentenças de citações pode vir a ser considerado um sumário de textos biomédico – sumarização focada em citação.

Além da TAC, a ST encontra espaço em várias outras conferências que se destinam a Linguística Computacional e ao PLN, assim como em oficinas e *workshops* de eventos. Os textos de língua inglesa são predominantes nas análises de pesquisas, mas há conferências que são voltadas para idiomas específicos, como o português na *International Conference on Computational Processing of Portuguese* (PROPOR) e no Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL).

Ambas as conferências, DUC/TAC, assim como a TREC (*Text Retrieval Conference*) criaram conjuntos de textos e normas que são consideradas referências à avaliação de métodos e sistemas de ST.

Há vários métodos e sistemas sumarizadores, porém os sistemas de notícias baseados na Web são os destaques. O Google News[9], o Columbia Newsblaster[10] e o News In Essence[11] extraem informação de múltiplos documentos na Web. Esses sistemas são capazes de identificar e lidar com a redundância dos documentos, além de reconhecerem novidades, o que assegura sumários completos e coerentes (DAS; MARTINS, 2007).

Outros sistemas que podem ser citados: NetSum – usa um algoritmo de redes neurais para melhorar as características das sentenças e gerar sumários extrativos a partir de um documento de entrada (SVORE; VANDERWENDE; BURGES, 2007). CSTSumm – é um sistema baseado na teoria discursiva CST (*Cross-document Structure Theory*) e destina-se a sumarização multidocumento em língua portuguesa (CASTRO JORGE; PARDO, 2010). GistSumm – é um sumarizador que usa a frequência das palavras para determinar a principal sentença do texto (*gist sentence*) e, a partir dessa, selecionar as sentenças de maior pontuação para construir sumários extrativos coerentes (PARDO; RINO; NUNES, 2003).

Para os próximos anos, como a informação está cada vez mais disponível na Web e em diferentes formatos, é provável que a sumarização multidocumento e *multi-lingual* será essencial. A mesma informação pode aparecer em vários documentos e em diferentes idiomas, portanto será necessário esforço para que seja apresentada em uma frase coerente e concatenada.

---

[9] https://news.google.com/
[10] http://newsblaster.cs.columbia.edu/
[11] http://newsinessence.com/

Por fim, o objetivo desta seção foi prover informações sobre o campo de pesquisa em ST, trazendo um breve histórico das abordagens mais conhecidas, mostrando as conferências internacionais a fim de destacar a atenção que a comunidade cientifica tem dado a área e, deixando claro que há espaço para evoluir.

A próxima seção apresenta a fundamentação sobre a lógica *fuzzy*.

2.2 FUNDAMENTOS DA LÓGICA FUZZY

Em 1965, Lotfi Asker Zadeh publicou o artigo *Fuzzy Sets* que apresentou a ideia de conjuntos com limites que não são precisos, recebendo o nome de lógica fuzzy (KLIR; YUAN, 1995). Zadeh criou a lógica "*fuzzy*" combinando os conceitos da lógica clássica e os conjuntos de Lukasiewicz (lógica dos conceitos "vagos").

Também chamada de lógica difusa ou nebulosa, a lógica *fuzzy* é considerada como uma extensão da lógica clássica e se diferencia desta devido a sua capacidade de se aproximar do mundo real. Além de lidar com a incerteza, a lógica *fuzzy* é capaz de modelar o raciocínio do senso comum, o que é difícil para os sistemas não probabilísticos.

A lógica clássica é limitada a dois valores – 0 (verdadeiro) ou 1 (falso); e isso facilita modelar o conhecimento em sistemas especialistas, por exemplo. Contudo, a limitação de valores da lógica clássica gera uma dedução de precisão (KYOOMARSI et al., 2008). Na teoria dos conjuntos *fuzzy*, cada elemento possuí um grau de pertinência restrito ao intervalo fechado de valores entre 0 e 1. Portanto, não existe uma fronteira bem definida para estabelecer quando um elemento pertence ou não a um determinado conjunto.

2.2.1 Conjuntos *crisp*

Na teoria de conjuntos clássicos (*crisp*), um conjunto é definido como uma coleção de objetos que compartilham certas características. Cada objeto individual é referido como um elemento ou membro do conjunto. Existem apenas dois graus de pertinência para definir se um elemento pertence ou não ao conjunto [0 ou 1]. Relacionamentos parciais não são concebidos.

Segundo Klir e Yuan (1995), considerando X o universo de discurso, existem três métodos básicos para definir um conjunto dentro do universo de X:

a) Nomeação de todos os elementos (método de lista). Este método pode ser usado apenas para conjuntos finitos, por exemplo: O conjunto A, onde seus elementos são $a_1, a_2, \dots, a_n$; pode ser escrito como: $A = \{a_1, a_2, \dots, a_n\}$.
b) Os elementos de um conjunto são definidos por uma propriedade. Uma notação para esse método é $A = \{x \mid P(x)\}$.
c) Um conjunto é definido por uma função (função característica), que declara quais os elementos de X são membros do conjunto ou não. Por exemplo: o conjunto A é definido pela função característica X_A da seguinte forma:

(11)

$$X_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

O elemento x é membro de A quando $\mathrm{X_A(x)} = 1$ ou não, quando $\mathrm{X_A(x)} = 0$.

Outra representação matemática desse relacionamento pode ser indicada pela função:

(12)

$$X_A: X \rightarrow [0\,,1]$$

2.2.2 Conjuntos *fuzzy*

Assim como os conjuntos *crisp* podem ser definidos por funções características, os conjuntos *fuzzy* podem ser caracterizados por funções de pertinência (*membership*). A função de pertinência μ associa a cada elemento x do conjunto A, de um universo X, um número real arbitrário $A(x)$, no intervalo fechado [0, 1], que caracteriza o grau de pertinência de x em A (KLIR; YUAN, 1995). A função de pertinência tem uma das seguintes formas:

(13)

$$\mu_A : X \rightarrow [0\,,1] \text{ ou } A : X \rightarrow [0\,,1]$$

Se o universo X é discreto e finito, com cardinalidade *n*, então o conjunto *fuzzy* A é representado na forma de um vetor de dimensão *n*. As entradas do vetor correspondem aos graus de pertinência dos elementos que compõem X. A notação de somatório pode ser utilizada para representar os elementos de X com grau de pertinência diferente de zero.

Por exemplo: se $X = \{x_1, x_2, \dots, x_n\}$, então o conjunto fuzzy $A = \{({}^{a_i}/_{x_i}) \mid x \in X\}$, onde $a_i = A(x_i)$ e $i = 1, \dots, n$ é dado por:

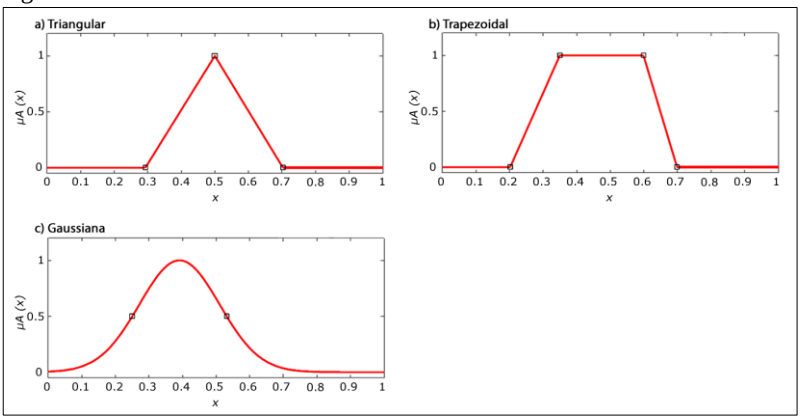
(14)

$$A = {}^{a_1}/_{x_1} + {}^{a_2}/_{x_2} + \dots {}^{a_n}/_{x_n} = \sum_{i=1}^{n} {}^{a_i}/_{x_i}$$

Nesta situação o símbolo Σ denota o conjunto de pares ordenados, não deve ser entendido com o somatório algébrico.

De acordo com Ibrahim (2004), as principais funções de pertinência são de forma triangular (a), trapezoidal (b) e gaussiana (c). A Figura 4 ilustra suas representações.

Figura 4 – SIF

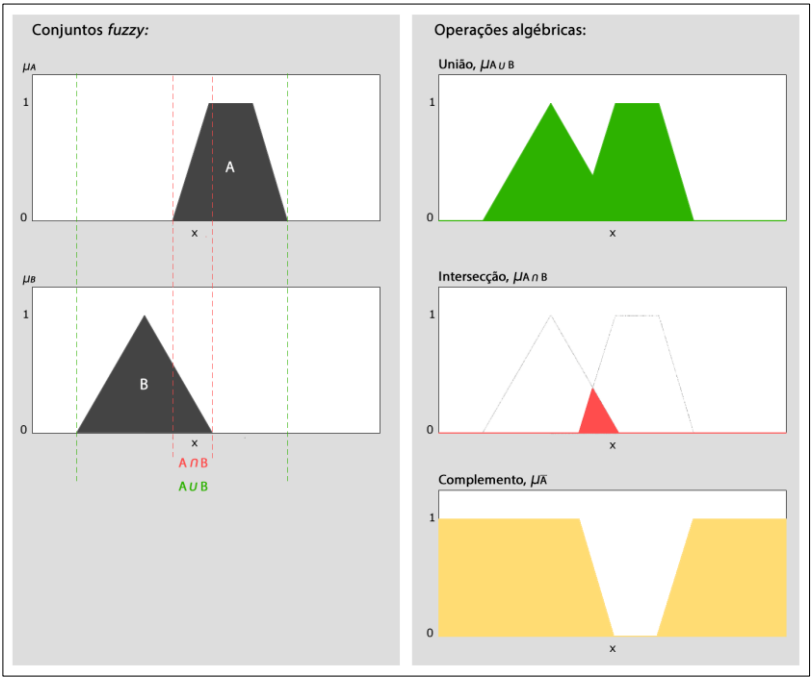


Fonte: Elaborada pelo autor

Cada função de pertinência representa um conjunto ou subconjunto *fuzzy*. O eixo das abcissas representa os números pertencentes ao conjunto e o eixo das ordenadas o grau de pertinência desses números ao conjunto. A leitura da notação $\mu_A(x) = 0{,}5$ diz que o grau de pertinência do número x ao conjunto A é de 0,5.

Conjuntos fuzzy podem ser manipulados algebricamente com operações da lógica clássica de união, intersecção e complemento definidas em termos do grau de pertinência dos conjuntos (IBRAHIM, 2004). Na Figura 5 as três operações básicas são ilustradas a partir de dois conjuntos *fuzzy*, A e B.

Figura 5 – Operações básicas



Fonte: Elaborada pelo autor

Para um elemento x com grau de pertinência $\mu_A(x)$ ao conjunto *fuzzy* A e $\mu_B(x)$ ao conjunto *fuzzy* B, a união e a intersecção dos conjuntos e o complemento de A são definidas pelas operações padrão:

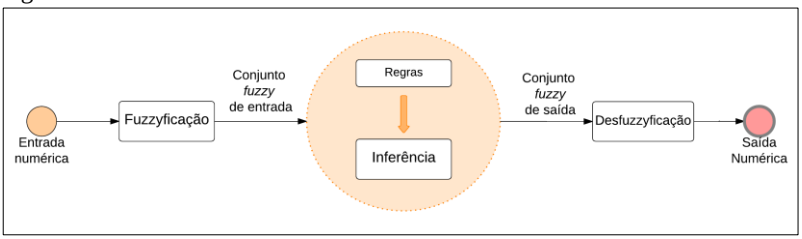
(15)

$$\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$$
$$\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$$
$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

2.2.3 Sistema de lógica fuzzy

Conforme o esquema da Figura 6, basicamente, são quatro os componentes que envolvem o Sistema de Lógica Fuzzy (SLF): Fuzzyficação, Inferência, Base de conhecimento (base de regras) e Desfuzzyficação.

Figura 6 – SLF



Fonte: Elaborada pelo autor

Um SLF processa uma entrada numérica (*crisp*) e gera uma saída também numérica. Para converter dados *crisp* em dados fuzzy, há um fuzzyficador no início do SLF e, para realizar o processo inverso, um desfuzzyficador ao final.

Existem duas abordagens para transformar dados numéricos em dados fuzzy, *singleton* ou *não-songleton*. A fuzzyficação *singleton* é usada para transformar um valor de entrada *crisp* em um *fuzzy singleton* (conjunto *fuzzy*), por meio de uma função de pertinência (ROSS, 2010). A fuzzyficação *não-singleton* é aquela cujas entradas são mapeadas para conjuntos *fuzzy* (triangular/gaussiana) que tem máximo grau de pertinência no valor da entrada *crisp* (MENDEL, 2007). A fuzzyficação *singleton* é a mais utilizada por ser mais simples e por requerer baixo custo computacional.

A Inferência é a fase em que se realiza o cálculo de todo o sistema *fuzzy*, conforme a base de conhecimento (regras). Ela determina como as regras são ativadas e combinadas através de operações entre as funções de pertinência. As operações de inferência mais comuns são o mínimo e o produto escalar.

As regras são obtidas por meio do conhecimento especialista ou dados numéricos. A base de conhecimento é composta por regras de estrutura condicionais do tipo *SE-ENTÃO*. O *SE* é a parte da condição, chamado de antecedente da regra e, o *ENTÃO* é resultado, chamado de consequente.

Existem vários Sistemas de Inferência Fuzzy (*Fuzzy Inference System* – SIF), mas os mais conhecidos são Mamdani-Assilian, Takagi-Sugeno-Kang (TSK) e Larsen (SCHMIDT; STEELE; DILLON, 2006). Os métodos Mamdani e TSK se diferem apenas na forma de obtenção da saída e no desempenho.

O SIF Mamdani é o mais conhecido e utilizado em aplicações devido sua estrutura simples e eficiente de operações de *min-max* (mínimo e máximo). Mamdani é adequado para aplicações analíticas

onde o conhecimento do especialista pode ser expresso sem um profundo conhecimento matemático. O SIF TSK é computacionalmente efetivo em situação onde a base de regras é grande e muitos ciclos de inferência são requeridos. Embora TSK seja computacionalmente eficiente, Mamdani é intuitivo, o que permite capturar o conhecimento especialista de forma intuitiva, semelhante à humana, tornando-se mais adequado para sistemas de sumarização de texto.

Por fim, a desfuzzyficação é onde se reduz o conjunto *fuzzy* para uma saída *crisp*, pois o resultado da inferência, que justamente serve de entrada para a desfuzzyficação, é um conjunto *fuzzy*. Um conjunto *fuzzy* não é entendido por sistemas convencionais, por isso é necessária à redução a um valor *crisp*.

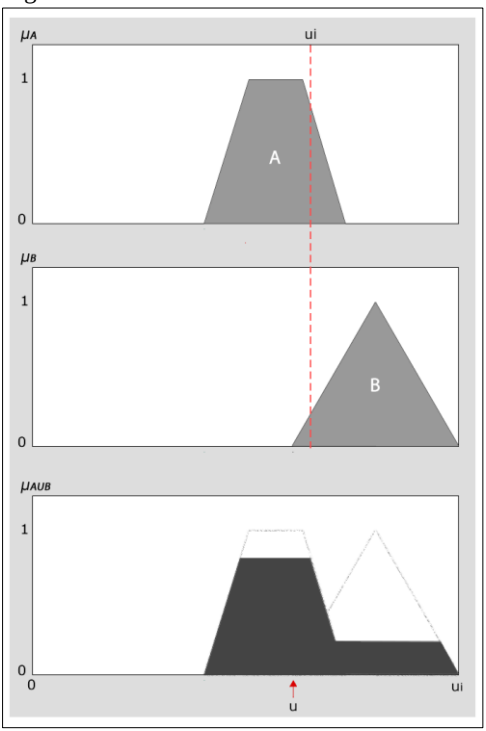
Há vários métodos de desfuzzyficação e, cada qual com vantagens e desvantagens. O método de desfuzzyficação centro de área, também chamado de centróide, é o mais conhecido e, no caso de valores discretos o resultado é obtido pela equação:

(16)

$$u = \frac{\sum_{i=1}^{N} u_i \mu_U(u_i)}{\sum_{i=1}^{N} \mu_U(u_i)}$$

Onde u é o índice correspondente ao centro de gravidade, μ_U é a união da área das funções de pertinência de saída e N é número de pontos de ativação. A Figura 7 mostra o método centroide quando um valor de entrada u_i ativa as funções de pertinências de A e B.

Figura 7 – Método centróide



Fonte: Elaborada pelo autor

No método *fuzzy* de sumarização proposto nesta dissertação, a desfuzzyficação tem a função de apresentar um valor de saída que corresponde à pontuação de uma sentença do texto.

2.3 ESTADO DA ARTE

Segundo Kitchenham (2004), uma revisão sistemática da literatura é o meio pelo qual se identifica, avalia e interpreta todas as pesquisas relevantes e disponíveis a um determinado assunto.

Para esse estudo, a revisão de literatura considerou artigos e livros publicados em bibliotecas digitais e base de dados, destacando-se: ACM Digital Library, IEEE Digital Library, Elsevier Digital Library, Scientific Digital Library (CiteSeer) Science Direct, Springer Link e Google Scholar. Também utilizou-se de trabalhos disponíveis em base

de teses e dissertações de universidades brasileiras, internacionais, além de repositórios de eventos e conferências.

Um estudo primário sobre a literatura do assunto conduziu a seleção dos termos de busca. Os termos de busca utilizados foram: *summarization*, *text summarization*, *automatic text summarization*, *extractive text summarization*, *significant words*, *sentence extraction*, *term frequency*, *fuzzy logic* e *fuzzy text summarization*.

O critério de inclusão dos trabalhos que fazem parte da bibliografia desse estudo levou em consideração o período da publicação; fator de impacto do periódico e conferência, no caso de artigos; apresentação de uma metodologia clara e resultados; a presença de uma bibliografia formal e o tópico abordado.

Após a leitura do *abstract* dos trabalhos e identificação dos critérios de inclusão, o artigo foi selecionado ou excluído. Os artigos selecionados, 127 ao todo, os quais 77 foram referenciados, sendo 13% do período de 2015 a 2013, 31% de 2012 a 2010, 24% de 2009 a 2007, 10% de 2006 a 2004, 9% de 2003 a 2001 e 13% em período inferior a 2000.

Existem diferentes abordagens para a SAT e inúmeros trabalhos publicados, porém, como o método proposto nesta pesquisa é embasado em lógica *fuzzy*, e técnicas comuns de pré-processamento de texto e recuperação da informação, para compor o estado da arte, foram considerados apenas os artigos relacionados aos tópicos mencionados e aderentes aos critérios de inclusão da revisão de literatura.

Nesta pesquisa buscou-se identificar o progresso do estado da arte dos procedimentos, métricas e algoritmos que tratam sobre a ST com lógica *fuzzy*, entre os anos de 2004 a 2014. Assim, levaram-se em conta os artigos que apresentaram propostas próximas a desta dissertação, e que são comentados na sequência.

No trabalho de Witte e Bergler (2007) foi proposto um algoritmo de clusterização para a geração de sumários a partir da detecção de tópicos comuns e distintos em um conjunto de documentos. O algoritmo é apoiado na teoria de conjuntos *fuzzy*, e segundo os autores, possui uma estrutura de dados flexível e de fácil adaptação para a filtragem de informações sensíveis ao contexto. Os conjuntos de dados da DUC de 2003 a 2006 foram selecionados para a avaliação dos sumários automáticos produzidos pelo algoritmo. Na avaliação utilizou-se a ferramenta ROUGE e os resultados foram comparados aos de outros sistemas. Os autores comentam, ainda, que o desempenho do algoritmo foi acima da média e muito próximo ao do melhor sistema, de 0,30 a 0,40, na média geral (medida-f).

Kyoomarsi et al. (2008) propuseram um método *fuzzy* de sumarização para gerar sumários de textos aplicados no TOEFL. Os autores comparam os resultados do método *fuzzy* proposto com um método de aprendizado de máquina (algoritmos C4.5 e Naïve Bayes). A avaliação consistiu no julgamento por cinco especialistas em língua estrangeira (professores de língua inglesa) que analisaram dez textos originais e seus respectivos sumários produzidos pelos métodos. Os juízes consideram o grau em que os principais conceitos do texto-fonte estavam presentes nos sumários automáticos de ambos os métodos. O método *fuzzy* apresentou resultados que superaram os do método baseado em aprendizado de máquina. Todos os juízes consideram os sumários do método *fuzzy* de melhor qualidade, tendo presente 77% dos principais conceito, em média, e 67% no método de aprendizado de máquina.

No trabalho de Suanmali, Binwahlan e Salim (2009a) o conjunto de dados DUC 2002 foi utilizado para testar o método *fuzzy* de sumarização de texto. Nove características serviram de entrada para o sistema de inferência, com funções de pertinências gaussianas. Em Suanmali, Salim e Binwahlan (2009b) o mesmo conjunto de testes foi utilizado, mas com oito características e funções de pertinência triangulares. Em ambos os casos, utilizou-se a ferramenta ROUGE para a avaliação, e os maiores valores de precisão, cobertura e medida-f foram obtidos pelo método *fuzzy*, 0,47 na média geral.

Leite e Rino (2009) descrevem um sumarizador extrativo em que a base de conhecimento *fuzzy* foi gerada por um algoritmo genético. Denominado de SuPor-2 Fuzzy, o sumarizador contempla onze métricas, cada qual modelada em três conjuntos fuzzy (baixo, médio e alto) por funções de pertinência triangulares. O corpus TeMário foi utilizado para o treinamento e, para a avaliação, ROUGE. A taxa de compressão dos sumários produzidos pelo SuPor-2 Fuzzy correspondeu a 30%, pois esse valor equivale-se a quantidade de informações nos sumários de referência. Na avaliação comparou-se o SuPor-2 a outros sistemas sumarizadores, e os melhores resultados foram obtidos pelo sumarizador proposto pelos autores, 0,74 e 0,73 respectivamente para ROUGE-1 e ROUGE-2.

Kiani e Akbarzadeh (2006) já haviam apresentado um algoritmo próximo ao da proposta de pesquisa de Leite e Rino (2009), inclusive com resultados semelhantes, mas sem uma avaliação automatizada.

No trabalho de Binwahlan, Salim e Suanmali (2010) apresentou-se um modelo híbrido de sumarização de texto, baseado em lógica *fuzzy*

e método do enxame de partículas, obtendo 0,45 na média geral com ROUGE-1.

Em Kyoomarsi et al. (2010) desenvolveu-se um método de sumarização de texto baseado em lógica *fuzzy* e na WordNet. Duas métricas foram definidas analisando-se os sinônimos das sentenças e de todo o texto. O método foi modelado com nove variáveis de entrada e quatro saídas. Os resultados com ROUGE e o conjunto de dados DUC 2003 alcançaram 0,60 na média geral.

Já em Suanmali, Salim e Binwahlan (2011) a abordagem sugerida para a ST combina lógica *fuzzy*, algoritmos genéticos e anotação de papel semântico. Os resultados com ROUGE apresentaram 0,47 na média geral.

Na pesquisa de Hannah, Geetha e Mukherjee (2011) descreveu-se um método *fuzzy* de sumarização de texto utilizando-se de sete características. As características foram categorizadas em três conjuntos *fuzzy* (menor, médio e maior) com funções de pertinência trapezoidais. O método *fuzzy* classifica as sentenças em não importante, importância média e importante. Os resultados com ROUGE e o conjunto de dados DUC 2002 mostraram um desempenho de 0,48 na média geral.

No trabalho de Megala, Kavitha e Marimuthu (2014) comparou-se o desempenho de dois métodos de sumarização de texto, em um deles utilizou-se a lógica *fuzzy*, no outro uma rede neural artificial. Os autores selecionaram textos de julgamentos jurídicos para os experimentos e procederam com uma avaliação não automatizada. O método sumarizador por redes neurais conseguiu 0,42 de desempenho na média geral, enquanto que o método *fuzzy* 0,46.

A discussão dos artigos possibilitou identificar quais as principais características textuais empregadas na sumarização extrativa, apresentadas ao final desta seção, e elaborar um quadro resumido (Quadro 2) dos trabalhos identificando os autores, o tipo de sumarização (monodocumento - MO ou multidocumento - MU), as características que compuseram as métricas (descritas no Quadro 3) e a avaliação. Todos os trabalhos relacionados no Quadro 2 empregam a técnica de extração.

Quadro 2 – Trabalhos de sumarização de texto com *fuzzy*

AUTOR (ANO)	VEICULAÇÃO	TIPO	DESCRIÇÃO* (Quantidade Características)	AVALIAÇÃO
Kiani e Akbarzadeh (2006)	IEEE International Conference on Fuzzy Systems	MU	b, 2c, d, i, g (5)	Precisão, cobertura e medida-f
Witte e Bergler (2007)	Advances in Artificial Intelligence	MU	a (1)	ROUGE**
Kyoomarsi et al. (2008)	IEEE/ACIS International Conference on Computer and Information Science Optimizing	MO	a, b, 2c, d, e, i, m, n, 2o (9)	Humano
Suanmali, Binwahlan e Salim (2009a)	Ninth International Conference on Hybrid Intelligent Systems	MO	2a, 2b, c, d, e, j, n (7)	ROUGE
Suanmali, Salim e Binwahlan (2009b)	International Journal of Computer Science and Information Security	MO	a, b, c, d, e, i, j, m (8)	ROUGE
Leite e Rino (2009)	Congresso da Sociedade Brasileira de Computação	MO	2a, 2c, d, e, 2i, 2n, p (7)	ROUGE
Binwahlan, Salim e Suanmali (2010)	Information Processing and Management	MO	2a, 3b, 3n (3)	ROUGE
Kyoomarsi et al. (2010)	Iranian Journal of Fuzzy Systems	MO	4a, 2c, 2d, p (4)	ROUGE
Suanmali, Salim e Binwahlan (2011)	IEEE International Conference on Dependable, Autonomic and Secure Computing	MO	a, b, c, d, e, i, j, m (8)	ROUGE
Hannah, Geetha e Mukherjee (2011)	Swarm, Evolutionary, and Memetic Computing	MO	a, b, d, e, i, j, m (7)	ROUGE
Megala, Kavitha e Marimuthu (2014)	International Journal of Computer Science and Information Technologies	MO	a, 3c, 2e, g, h, i, m (7)	Precisão, cobertura e medida-f

Fonte: Elaborado pelo autor
*a descrição das características encontra-se no Quadro 3.
**Avaliação automatizada.

Com exceção do trabalho de Witte e Bergler (2007), a principal crítica sobre os trabalhos no Quadro 2 é com relação a quantidade de

métricas empregadas nos métodos. Por exemplo, em Suanmali, Binwahlan e Salim (2009a) foram selecionadas sete características sendo derivadas em onze métricas. Como mencionado no Capítulo 1 desta dissertação, definir quais as melhores métricas para alcançar o conteúdo informativo de documentos não é algo simples, pois fatores como a estruturação do texto e a linguagem utilizada podem influenciar no desempenho das métricas.

Quando utiliza-se a lógica *fuzzy* para sumarizar texto, a quantidade de métricas pode tornar-se um problema, já que quanto maior o número, maior será a quantidade de regras para expressar o conhecimento. Além disso, aumentando-se o número de variáveis linguísticas no antecedente da regra, aumenta-se a complexidade do problema, o que pode comprometer o desempenho do sistema.

Portanto, o método *fuzzy* de sumarização de texto proposto nesta pesquisa traz a ideia da redução da dimensionalidade do problema por meio da correlação de características. Duas características textuais distintas foram utilizadas para compor uma nova métrica e implementada no FSumm.

As características textuais, assim como a abordagem ou o método empregado no processo de ST são os responsáveis pelo resultado, o sumário. Para Nomoto e Matsumoto (2003), um sumário é considerado adequado quando este representa todo o conteúdo do documento(s).

Todas as características estatísticas e linguísticas descritas nessa seção são importantes para o processo de sumarização de texto, porém o resultado desse processo depende de como elas são exploradas e combinadas.

Quadro 3 – Descrição das características do texto

ID	CARACTERÍSTICAS	DESCRIÇÃO
a	Frequência das palavras	Parte do pressuposto de que a ideia principal de um texto pode ser expressa pela representatividade da frequência das palavras, utilizando-se de métodos estatísticos ou que envolva análise morfológica.
b	Título	Palavras do título ou cabeçalho são indicativos do tema do documento.
c	Posição	Pode envolver o posicionamento da sentença com relação ao parágrafo, com relação a uma seção do texto, ou ainda, com relação ao documento todo. A primeira e a última sentença de um parágrafo tendem a ser mais importante, assim como o primeiro e o último parágrafo.
d	Comprimento	Refere-se a quantidade de palavras da sentença.
e	Substantivos próprios	Palavras que são nomes de pessoas, lugares, entidades nomeadas ou que representam um conceito.
f	Acrônimos	Siglas também apresentam importância para o peso de uma sentença, pois podem ser nomes.
g	Frase sugestiva	São sentenças com frases sugestivas, tais como: “em conclusão, este artigo, este relatório, desenvolver, o objetivo, entre outras”.
h	Palavras influentes	Uma lista de palavras sobre um domínio específico pode ser definida.
i	Palavras temáticas	São palavras que expressam tópicos discutidos no documento.
j	Informação numérica	Sentenças com informações numéricas são importantes.
k	Estilo das palavras	Palavras com grifo (negrito, itálico e sublinhado), ou ainda, em maiúsculo, também são importantes.
l	Pronomes	Pronomes não são incluídos em características importantes, a menos que estejam acompanhados de um substantivo correspondente.
m	Coesão sentença-sentença	Calcula-se a semelhança de uma sentença com as demais e adiciona-se um valor de similaridade.
n	Coesão sentença-centróide	Para cada sentença s se computada o vetor representando o centróide do documento, que é a média aritmética sobre os valores das coordenadas correspondentes de todas as sentenças do documento. Após isso, calcula-se a similaridade entre o centroide e cada uma das sentenças, obtendo um valor bruto para cada sentença.
o	Informações não essenciais	São palavras que ocorrem, muitas vezes, no início de frases, tais como: “porque, devido, além disso, geralmente, normalmente.” Está é uma característica binária, onde a sentença assume o valor verdadeiro se apresentar pelo menos um desses marcadores de discurso, e falso caso contrário.
p	Análise de discurso	Informação sobre o nível do discurso em um texto também é uma boa característica. Analisando-se o discurso é possível identificar a estrutura geral do texto e, em seguida, remover frases que são periféricas até que se obtenha a principal mensagem.

Fonte: Elaborado pelo autor

3 PROPOSTA DO MÉTODO FUZZY

A proposta desta dissertação é um método *fuzzy* de sumarização extrativa com base nas características estatísticas e linguísticas do texto, denominado de FSumm. Por meio desse método, uma sentença é representada como um vetor de termos de onde são extraídas as informações estatísticas. As informações estatísticas são inferidas pelo sistema *fuzzy* com o objetivo de classificar e identificar quais são as principais sentenças do texto original.

A relevância textual é definida pela extração de métricas provenientes das abordagens de frequência das palavras, posição e comprimento das sentenças.

Para a frequência das palavras foi selecionada a métrica *tf-isf* e, após a análise preliminar dos dados gerados a partir do conjunto de textos para a avaliação, as outras métricas foram sugeridas. A métrica de posicionamento foi aprimorada, a de comprimento foi elaborada uma nova e, a correlação entre a posição e o comprimento foi definida por meio de regressão linear multivariada que converge as duas métricas anteriormente citadas em uma única medida, chamada de *loc-len* (do inglês *local* e *length*, posição e comprimento). A métrica *loc-len* foi validada e utilizada como variável de entrada no método *fuzzy*, substituindo a posição e o comprimento como entradas distintas.

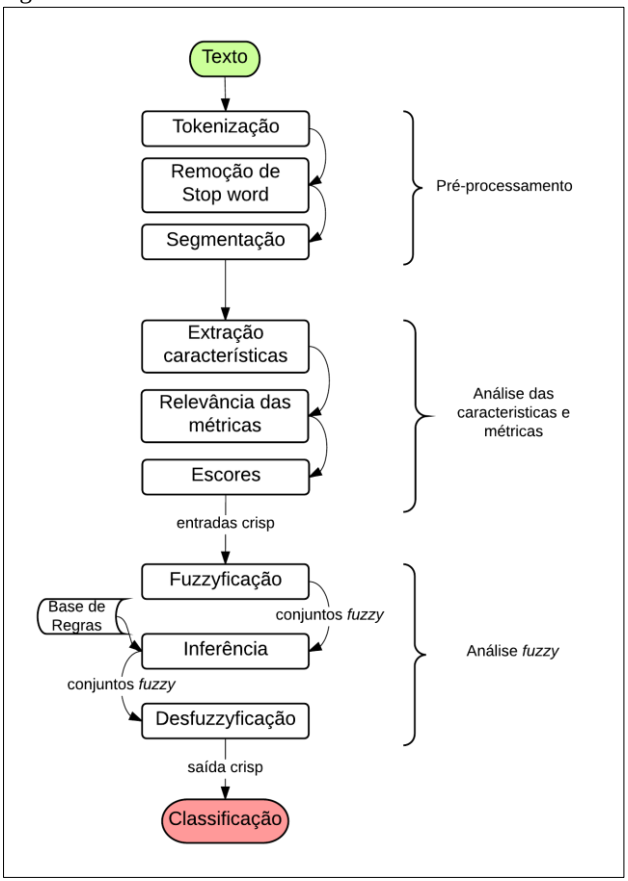
Para a construção de um extrato pela justaposição das unidades de texto selecionadas, diversas técnicas possibilitam ordenar as sentenças, desde as mais simples, que utilizam informações básicas, até as mais complexas, as quais fazem uso de análise semântica das relações. Lima e Pardo (2012) analisaram vários métodos de ordenação de sentenças para sumários multidocumentos e constataram que a posição das sentenças dos sumários ordenados manualmente tem alta correlação com a posição das sentenças no texto-fonte, ao que se refere às posições inicias. Embora o método baseado na posição das sentenças seja simples, foi o mais eficaz. Por esse motivo, o FSumm segue a ordenação baseada na posição.

O propósito do FSumm é a produção automática de extratos informativos e genéricos de um texto-fonte, empregando-se uma abordagem superficial (empírica) por meio de conceitos *fuzzy* e métricas já conhecidas ou, até então novas, em processamento de linguagem. Portanto, o método *fuzzy* de sumarização desenvolvido não emprega *parsers* ou modelos de discurso, mas alia-se a sumarização extrativa a técnicas e ao modelo vetorial de RI.

3.1 O MÉTODO PROPOSTO

A estrutura geral do FSumm é composta de três processos-chaves: pré-processamento, análise das características textuais e análise difusa. A Figura 8 ilustra o fluxo do método e suas etapas.

Figura 8 – Estrutura do método



Fonte: Elaborada pelo autor

O primeiro processo envolve a preparação do texto para a análise. O processo seguinte é onde ocorre o cálculo estatístico das características. A análise *fuzzy* é o processo onde os analisadores difusos entram em cena para classificar as sentenças com base na estatística da etapa anterior. Por fim, as sentenças são selecionadas de acordo com a classificação.

Alguns métodos de sumarização, anteriormente propostos, lidavam com as características textuais como parâmetros binários. Atribuía-se valor 0 ou 1, que não funcionam com precisão o tempo todo (KYOOMARSI et al., 2010).

Uma forma de contornar o problema de parâmetros binários, referente a dados *crisp*, é utilizar qualidades difusas. Isto significa que cada sentença terá um valor de relevância compreendido entre os reais existentes no intervalo fechado [0, 1], dependendo das características específicas que apresentar.

3.2 DEFINIÇÃO DO PROBLEMA

Assumindo-se que a tarefa de sumarização é encontrar um subconjunto de sentenças, que de alguma forma representam o conteúdo principal do texto-fonte, segue uma formalização matemática para o problema.

Um $Texto$ pode ser representado da seguinte forma:

$$Texto = \{ < a_n, t_i, s_m > | \; a \in \Sigma; t \in T \cup \Sigma; \; s \in B_y \subset S \}$$

Sendo composta de:

a) Uma tripla formada por a, t e s denominados de símbolos do alfabeto, termos (palavras) e sentenças (frases), respectivamente;
b) Um conjunto Σ dos símbolos que definem o alfabeto de uma linguagem, onde $\Sigma = \{a_1, a_2, \dots, a_n\}$;
c) Um conjunto T dos termos do documento, onde $T = \{t_1, t_2, \dots, t_i\}$. Um termo pertence ao conjunto T ou Σ (apenas um símbolo). A combinação de símbolos do alfabeto, Σ^* (operação estrela de Klenne), forma um termo;
d) Um conjunto S contendo as sentenças dos documentos, onde $S = \{s_1, s_2, \dots, s_m\}$, e um conjunto B_y contendo um grupo de sentenças de S, portanto $B_y \subset S$;

3.3 IMPLEMENTAÇÃO DO FSUMM

Uma das atividades previstas nesta dissertação é testar ferramentas que poderiam auxiliar no pré-processamento. Foram analisadas às ferramentas Rapidminer, Weka e Nltk. Todas bastante utilizadas em mineração de texto e PLN. A ferramenta que se demonstrou mais promissora para o contexto da pesquisa foi a Nltk, mas com ressalvas se o texto processado estiver em língua portuguesa.

Para processar texto na ferramenta Weka os arquivos precisam ser transformados para uma extensão que o software compreenda, além disso, é preciso saber dentre os vários filtros de atributos quais os que podem ser utilizados. A Nltk precisa que os textos em português sejam decodificados para a análise, e quando necessário exportar pode ocorrer perda de informação estrutural.

A maior dificuldade encontrada com as ferramentas no pré-processamento foi no momento da exportação das informações. Os textos exportados precisam ser lidos e analisados pelo FSumm, mas com a exportação houve perda de informação estrutural, o que poderia comprometer o resultado das métricas. Por esse motivo, optou-se por desenvolver uma aplicação com todas as etapas necessárias para o método.

O FSumm foi implementado com o CodeIgniter[12], um *framework* para desenvolvimento de aplicações na linguagem de programação PHP. O CodeIgniter usa a abordagem MVC (*Model-View-Controller*) a qual permite a separação entre o que está na camada lógica e de apresentação, possui um abrangente conjunto de bibliotecas voltadas as tarefas mais comuns e com bom desempenho (leve e rápido).

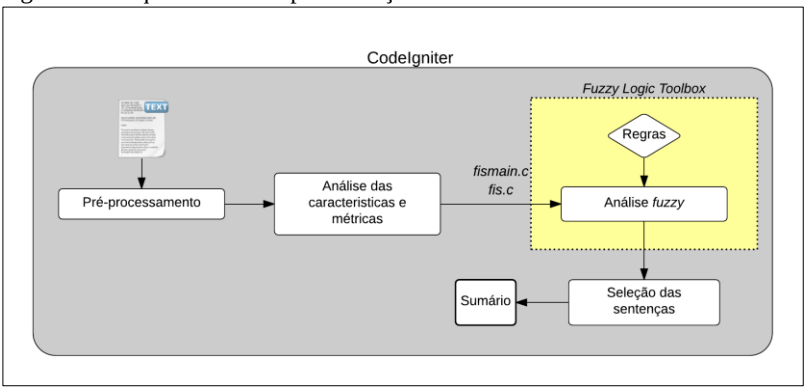
Durante o desenvolvimento da aplicação, para fins de utilização efetiva, pesquisou-se por bibliotecas e ferramentas que aplicassem a lógica *fuzzy*. Os aspectos desejáveis a seleção da ferramenta foram o baixo custo, disponibilidade para download e a possibilidade de integração. Optou-se por utilizar o SIF implementado em linguagem C do Matlab, presente no *Fuzzy Logic Toolbox*. O Matlab é um software interativo de alta performance voltado para o cálculo numérico.

O *Fuzzy Logic Toolbox* é um *software* com uma interface gráfica bastante intuitiva para a análise, construção e simulação de sistemas baseados em lógica *fuzzy*. O *toolbox* é capaz de gerar motores de inferência *fuzzy* em código C incorporável ou *stand-alone* executáveis. Existem dois arquivos em código C disponibilizados pelo *toolbox*,

[12] http://www.codeigniter.com/

fismain.c e *fis.c*, que são o código-fonte do motor de inferência *fuzzy* no modo *stand-alone*. O código-fonte permite a aplicação ler os arquivos *fuzzy* (de extensão FIS) criados no Matlab. A Figura 9 ilustra a arquitetura da implementação do método.

Figura 9 – Arquitetura da implementação do método



Fonte: Elaborada pelo autor

No Quadro 4 está incluído um trecho do texto que servirá de referência às seções seguintes, com o propósito de ilustrar o que ocorre em cada processo-chave do método. No texto está destacado o título [T], parágrafos [P] e sentenças [s].

Quadro 4 – Texto do caderno Opinião do jornal Folha de São Paulo, presente no corpus TeMário.

[T] Como direcionar sua empresa para o cliente.

[P1][s1] É preciso inverter as estruturas para diminuir a distância entre o cliente e os que detêm o poder de decisão.

[P2][s2] Empresa fadada ao insucesso tem duas caras: uma real, outra para o cliente. **[s3]** Na hora de vender, promessas; quando o cliente confere, decepções. **[s4]** É difícil encontrar o responsável, quando a empresa não é direcionada à satisfação total dos clientes.

[P3][s5] Nas estruturas tradicionais de empresas, onde o mando predomina sobre a responsabilidade individual, não é possível sequer aprender em cima dos próprios erros. **[s6]** Faz-se de tudo para que não haja registro do erro, para que ele não seja do conhecimento dos que detêm o poder de mando.

...

[P11][s7] A responsabilidade compartilhada e o trabalho em equipe só poderão se desenvolver se a estrutura permitir uma interação constante entre as áreas.

[P12][s8] A empresa toda é um macroprocesso, uma equipe única voltada para o objetivo comum de atingir altos níveis de produtividade, com a manutenção e a conquista de novos clientes.

...

[P15][s9] A perda de competitividade é a consequência mais direta da falta de agilidade nas decisões, perda de informações, aumento da burocracia interna, pois, nos "momentos da verdade", o funcionário precisa tomar decisões que implicam, muitas vezes, questões vitais para o cliente. **[s10]** Frequentes consultas aos níveis superiores causam perda de tempo e dinheiro. **[s11]** A redução dos níveis hierárquicos ao mínimo necessário traz agilidade. **[s12]** Experimente.

Fonte: (PARDO; RINO, 2003).

3.4 PRÉ-PROCESSAMENTO

Métodos padrão de pré-processamento de RI e mineração de texto são aplicados. A tokenização é utilizada para decompor o texto de entrada em cada palavra que o compõe. *Stop words*, palavras de significado irrelevante (podem ser: artigo, preposição, advérbios, números, pronomes e pontuação) são removidas das sentenças. A segmentação consiste em delimitar o texto de entrada em sentenças.

Uma sentença é um conjunto de palavras que formam uma oração. A primeira palavra deve iniciar com letra maiúscula, e a ultima ter o símbolo subsequente de final de oração (ponto final, interrogação ou exclamação). Ao final, cada sentença estará representada na forma de um vetor de termos.

O resultado do pré-processamento para o texto do Quadro 4 são quarenta sentenças, tais como as sentenças $s2$, $s3$ e $s4$ do parágrafo $P2$, aqui sendo colocadas como exemplo:

$$s2: [empresa, fadada, insucesso, duas, caras, real, cliente]$$
$$s3: [hora, vender, promessas, cliente, confere, decepções]$$
$$s4: \begin{bmatrix} difícil, encontrar, responsável, empresa, direcionada, \\ satisfação, total, clientes \end{bmatrix}$$

O texto original contém 614 palavras e o texto pré-processado contém 302 palavras, levando a uma redução de aproximadamente 49% do tamanho do texto original.

Todas as etapas do pré-processamento de um texto contemplam as línguas inglesa e portuguesa (do Brasil), a fim de possibilitar uma avaliação bilíngüe do sistema.

3.5 ANÁLISE DAS CARACTERÍSTICAS E MÉTRICAS

O cálculo das características é gerado de forma local e global. Local quando envolvem critérios de localização, quantificação dos termos nas sentenças e no texto como um todo; e global, através da definição da similaridade entre os termos das sentenças. As características selecionadas para este estudo foram:

- **Posição**: é a localização das sentenças no texto. Está característica pode envolver o posicionamento da sentença com relação ao parágrafo, com relação a uma seção do texto, ou ainda, com relação ao documento todo. A primeira e a última sentença de um parágrafo tendem a ser mais importantes, assim como o primeiro e o último parágrafo.
- **Comprimento**: é a quantidade de termos presentes na sentença. Sentença muito longa ou curta, geralmente não é incluída no sumário.
- **Título**: palavras do título são indicativos do tema do documento, sendo assim, sentenças com palavras do título são mais importantes do que as sentenças onde isso não ocorre.
- **Keywords**: é a frequência dos termos na sentença e no texto. *Keywords* são nomes de coisas. Para determinar as *keywords*, pode-se utilizar a medida *tf-idf* ou algum método que envolva análise morfológica. Partindo do pressuposto de que a ideia principal de um texto pode ser expressa por um conjunto de *keywords*, a sentença que apresentar *keywords* têm maior chance de ser incluída no sumário.

O motivo pela escolha dessas características deve-se ao fato de serem as comumente encontradas nas pesquisas em SAT extrativa. Os

critérios das características provêm das Seções 2.1.2.1 a 2.1.2.3 e do Capítulo 3 desta dissertação e dos trabalhos de Brandow, Mitze e Rau (1995); Radev, Hovy e McKeown (2002); Gupta e Lehal (2010); Kyoomarsi et al. (2010); Alguliev et al. (2011); Mei e Chen (2011); Suanmali, Salim e Binwahlan (2011); Kiabod, Dehkordi e Sharafi (2012), Aries, Oufaida e Nouali (2013) e Camargo (2013).

3.5.1 Posição

Para encontrar a posição das sentenças no parágrafo, conforme sua importância, o cálculo P_0 se define pela razão entre a posição de cada sentença n_i e o número de sentenças do parágrafo n_s.

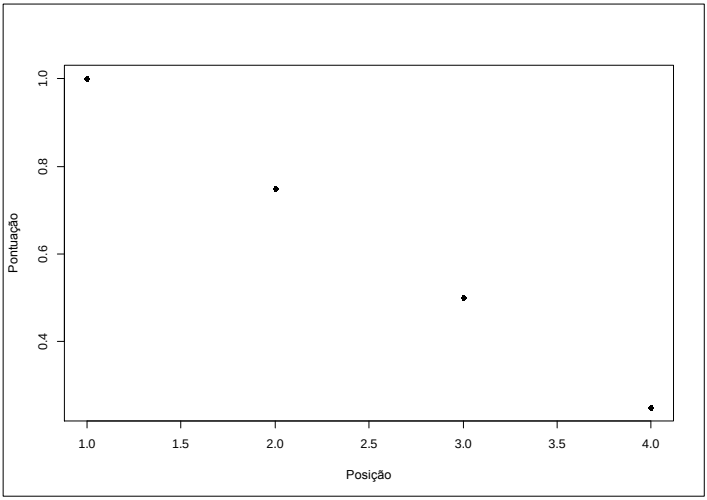
(17)

$$P_0 = \frac{p_i}{n_s}, \qquad onde\ p_i = n_s\ e\ p_i = n_s - 1, \dots, 1$$

O P_0 é um valor normalizado entre 1 e 0. A sentença na posição mais significativa terá o $P_0 = 1$ enquanto que a última terá P_0 próximo de 0. A Equação para P_0 é uma maneira simples de ranquear sentença conforme sua posição no texto ou no parágrafo e também é a forma mais encontrada nos trabalhos que utilizam essa métrica.

No parágrafo $P15$ do Quadro 4 há quatro sentenças: $s9$, $s10$, $s11$ e $s12$. A pontuação de cada uma delas no parágrafo pode ser dada por $s9 = \frac{4}{4} = 1$; $s10 = \frac{3}{4} = 0{,}75$; $s11 = \frac{2}{4} = 0{,}5$; $s12 = \frac{1}{4} = 0{,}25$. A Figura 10 demonstra o comportamento da medida para as quatro sentenças do parágrafo utilizado como exemplo.

Figura 10 – Pontuação da posição do parágrafo $P15$.



Fonte: Elaborada pelo autor

A relação entre a pontuação (peso) e a posição da sentença no parágrafo é linear. Em decorrência do comportamento linear que a medida P_0 assume, não é possível confirmar a justificativa encontrada na literatura, pois apenas as sentenças iniciais são valoradas com peso maior, o que não acontece com as últimas. A P_0 também pode ser empregada para pontuar apenas os parágrafos.

Com relação a quantidade de sentenças no parágrafo, dependendo da forma como o texto está estruturado, vários parágrafos podem conter apenas uma ou duas sentenças, em ambos os casos as sentenças receberiam a pontuação máxima. Portanto, diante das questões expostas sobre a medida P_0, uma nova medida de posição foi elaborada, denominada de loc, sendo estimada pela função:

(18)

$$loc(p_{iy}, n_{Sy}) = \begin{cases} 1 - \left(\dfrac{p_{iy} - 1}{n_{Sy}}\right) & ; \ Se \ y = 1 \\ \dfrac{p_{iy}}{n_{Sy}} & ; \quad Se \ y = 2 \end{cases} \qquad onde \ p_{iy} = 1, \dots, n_{Sy}$$

O texto é dividido em dois blocos de sentenças B_1 (superior) e B_2 (inferior), os quais são analisados individualmente. Sendo S o conjunto das sentenças s do texto, as sentenças pertencem a um dos subconjuntos

de B_y. O cálculo loc é definido pela razão entre a posição de cada sentença p_{iy} no texto e o número de sentenças n_{Sy} no subconjunto B_y.

Aplicando loc no texto do Quadro 4, as sentenças $\{s1; \dots; s6\} \in B_1$ e $\{s7; \dots; s12\} \in B_2$, com uma pontuação $\{1; \dots; 0{,}16\} \in B_1$ e $\{0{,}16; \dots; 1\} \in B_2$, respectivamente.

A nova métrica da posição pontua a primeira sentença de B_1 com o valor máximo e a ultima com o valor mínimo, no caso das sentenças de B_2 acontece o inverso.

3.5.2 Comprimento

Uma forma de calcular a relevância de uma sentença em relação ao seu comprimento é fazendo a razão da quantidade de termos pela quantidade de termos da maior sentença do texto (DAS; MARTINS, 2007). Dependendo da estruturação do texto, sentenças demasiadamente longas podem ser privilegiadas.

No trabalho de (RINO; PARDO, 2003) o tamanho das sentenças foi definido pela média dos termos. Uma escala de quatro patamares, variando de 0 até o valor da maior sentença, categoriza o tamanho das sentenças. A média dos termos está sendo utilizada como um parâmetro de delimitação do tamanho da sentença. Não há uma métrica que estabeleça um valor de corte sobre o tamanho ideal da sentença que deve compor o sumário, pois cada texto possui formas distintas.

Esses fatos motivaram a formulação da métrica len que é definida da seguinte forma:

$$len = ln\left(\bar{t}_{Sy} - \left|\frac{\bar{t}_{Sy} - t_{siy}}{\sigma_y}\right|\right) \; onde \; y = 1, 2 \tag{19}$$

O valor de len é calculado pelo logaritmo natural da média da quantidade dos termos $\bar{t}_{Sy}$ (no subconjunto B_y), menos o valor absoluto do escore padrão. Onde t_{siy} é a média da quantidade dos termos da sentença observada e σ_y o desvio padrão. O peso que resulta da len aumenta à medida que a quantidade de termos da sentença se aproxima da média $\bar{t}_{Sy}$.

Diante dessa especificação, o valor do atributo len de uma dada sentença é calculado levando-se em conta apenas as palavras descritivas, pois denotam conteúdo semântico. Por exemplo, no Quadro 4, a $s2$ que contém treze palavras passará a conter sete, como demonstrado na Seção 5.1. A média dos termos em B_1 é $\bar{t}_{S1} = 7{,}8$, dessa forma, $s2$ terá o

atributo $len = 1{,}99$; já a $s3$, por conter menos palavras descritivas terá $len = 1{,}91$.

3.5.3 Correlação entre a posição e o comprimento

A $loc\text{-}len$ é uma métrica que converge os atributos posição e comprimento para uma pontuação comum. A forma da equação é a seguinte:

$$loc\text{-}len = loc \times len \tag{20}$$

O produto entre a loc e len pontua uma sentença considerando a posição desta no texto e a quantidade de termos, tendo valor crescente a partir da região central, em direção às extremidades.

Para estimar $loc\text{-}len$ em função de Posição loc e do Comprimento len, é possível conceber um modelo por meio da técnica de regressão. A regressão trata da questão de estimar um valor condicional esperado, onde o método de estimação amplamente utilizado é o método dos Mínimos Quadrados Ordinários (MQO) (BARBETTA; REIS; BORNIA, 2010).

O modelo que permite descrever a relação entre mais de duas variáveis quantitativas é chamado de Regressão Multivariada. A relação entre a variável dependente (variável de resposta) com as variáveis independentes (exploratórias) é dada por uma equação matemática, estruturada na seguinte forma:

$$Y = f(x_1, x_2, \dots, x_k) + \varepsilon \tag{21}$$

Onde Y é a variável de resposta, x_1 e x_2 são as variáveis exploratórias com x_i sendo a i-ésima variável exploratória e $i = 1, \dots, k$ e, ε é o erro aleatório. Sendo assim, $loc\text{-}len$ é a variável de resposta e t_{siy} (quantidade de termos) e loc são as variáveis exploratórias.

Uma vez estimado o modelo de regressão multivariada, pode-se predizer o valor da variável $loc\text{-}len$ a partir dos valores do conjunto de variáveis t_{siy} e loc. O modelo de regressão estimado é apresentado na Seção 6.3.1, já que faz parte dos experimentos. A ideia de conceber uma métrica, com base em um modelo que correlaciona características textuais distintas é algo ainda não explorado na ST.

3.5.4 Keywords

Ao se quantificar os termos do documento, é possível usar a sua frequência para se atribuir um peso. O número de vezes que um termo aparece no documento aumenta proporcionalmente o seu peso, sendo inversamente proporcional a frequência do termo no documento. Segundo Baeza-Yates e Ribeiro-Neto (2011), essa afirmação é baseada na observação de que a alta frequência dos termos é importante para descrever documentos.

Um dos critérios de similaridade mais utilizados em tarefas de RI é a medida *tf-idf*, e com base nela, a medida *tf-isf* (frequência do termo – frequência inversa da sentença) calcula o valor médio da frequência de todos os termos do texto. Dessa forma, define-se um escore à sentença a partir da frequência dos termos:

(22)

$$w_{ij} = tf_{ij} \times isf = \frac{f_{ij}}{\sum_{k=1}^{N} f_{kj}} \times log \frac{N}{|n_i|}$$

O peso w_{ij} é dado pela frequência tf_{ij} multiplicada pela medida da importância geral do termo isf. A f_{ij} é a quantidade de ocorrências do termo t_i no documento d_j e k é a quantidade de termos distintos, sendo $k \in T$. Entretanto, é comum a frequência ser normalizada para prevenir um viés em documentos longos ou quando se lida com mais de um documento. A isf é definida como sendo o logaritmo do quociente entre o número total de sentenças do documento N e o número de sentenças que contem o termo n_i. O escore K_1 é definido pela razão do somatório dos pesos dos termos w_{ij} da sentença s_i com o maior somatório dos pesos dos termos das sentenças s_i^N.

(23)

$$K_1 = \frac{\sum_{i=1}^{k} w_{ij}(s_i)}{Max(\sum_{i=1}^{k} w_{ij}(s_i^N))}$$

Uma sentença pode conter palavras que se assemelham ao título ou subtítulo. Cada termo do título é consultado a cada sentença, estabelecendo assim um modelo espaço vetorial ponderado.

(24)

$$K_2 = sim(s_i, q_y) = \sum_{j=1}^{k} w_{ij}\, w_{yj}$$

Onde s_i é a i-ésima sentença, q_y é a consulta do título, w_{ij} é o peso do termo da sentença e w_{yj} é o peso do termo do título.

Todas as métricas apresentadas são normalizadas para melhorar o processo de análise e avaliação. O Quadro 5 lista as medidas utilizadas na implementação do método, uma descrição e o critério científico.

Quadro 5 – Características textuais

CARACTERÍSTICA	VARIÁVEL	DESCRIÇÃO	CRITÉRIO
Posição e Comprimento	$loc\text{-}len$	Correlação da posição e do comprimento das sentenças no texto.	As primeiras e as últimas sentenças do texto tendem a ser mais importantes, assim como as que apresentam a quantidade de termos próxima da média dos termos de todas as sentenças.
Keywords	K_1	O número de termos da sentença (*tf-isf*).	A sentença que apresentar *keywords* tem maior chance de ser incluída no sumário.
	K_2	O número de termos do título na sentença.	Sentenças com palavras do título são mais importantes do que onde isso não ocorre.

Fonte: Elaborado pelo autor

3.6 ANÁLISE FUZZY

Segundo Sivanandam, Sumathi e Deepa (2007), seis passos devem ser seguidos para calcular a saída:

a) Determinar o conjunto de regras *fuzzy*;
b) Fuzzyficar as entradas usando as funções de pertinências de entrada;
c) Combinar as entradas fuzzyficadas de acordo com as regras *fuzzy* para estabelecer a força da regra;
d) Encontrar o consequente da regra para combinar a regra forte com a função de pertinência de saída;
e) Combinar os consequentes para obter uma distribuição de saída;
f) Desfuzzyficar a distribuição de saída (apenas se necessário uma saída *crisp*).

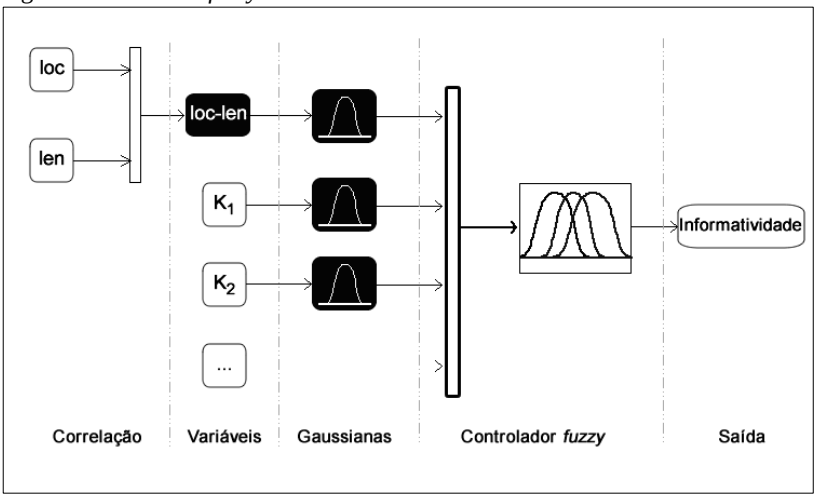
Os parâmetros textuais extraídos são utilizados como entrada para o SIF. O universo de discurso para as variáveis de entrada foi

particionado em vários conjuntos *fuzzy* designados por funções gaussianas. O motivo para a escolha de funções de pertinência gaussianas é devido a curva ser suave e diferente de zero em todos os pontos.

Conjuntos *fuzzy* na forma triangular/trapezoidal são compostos por funções lineares com equações simples. Isso torna a computação necessária para as operações matemáticas sobre as regras mais fácil do que para outros tipos de funções, porém, pode apresentar valores desfuzzyficados repetidos.

Mamdani e centro de gravidade foram utilizados nesta pesquisa para realizar a análise *fuzzy* em um processo de agregação sobre as características, com o operador de implicação "mínimo". São três as variáveis de entrada do SIF: posição e comprimento (*loc-len*) e *Keywords* (K_1 e K_2), conforme demonstra a Figura 11.

Figura 11 – Modelo *fuzzy*



Fonte: Elaborada pelo autor

3.6.1 Modelagem *fuzzy*

A ideia da modelagem *fuzzy* é a de descrever o comportamento do sistema para sua análise, simulação e projeto. Dentre suas vantagens, destaca-se a capacidade para descrever sistemas complexos que lidam com o conhecimento e o uso de variáveis linguísticas que facilitam a compreensão do problema.

No caso do FSumm, as entradas na modelagem são os escores gerados no processo de análise das características. As três variáveis de entrada *loc-len*, K_1, K_2 e a variável de saída *Informatividade* são particionadas em conjuntos *fuzzy*. A cada conjunto *fuzzy* definido sobre um universo comum, associa-se um termo linguístico e uma função de pertinência. Por exemplo, a variável *loc-len* assume como valor um dos membros do conjunto {Baixo, Médio, Alto}.

O sistema é representado por regras *fuzzy* que são responsáveis por expressar o conhecimento a partir do conjunto de termos linguísticos das variáveis de entrada e saída.

Regras *fuzzy* são declarações condicionais que envolvem variáveis linguísticas. Segundo Schmidt, Steele e Dillon (2006), a variável linguística expressa em linguagem natural o nome do conjunto *fuzzy* composto de possíveis valores numéricos que uma quantidade de interesse pode assumir. Foram definidas vinte e sete regras na base de regras do FSumm, onde i de R_i varia de $i = 1, \dots, 27$, baseadas nos trabalhos de Zadeh (1999) e Brandow, Mitze e Rau (1995). As regras são estruturadas da seguinte forma na base de regras:

R_1:

$$\textbf{\textit{Se}}\ (\textit{loc-len}\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_1\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_2\ \text{é}\ \textit{Baixo})\ \textbf{\textit{Entã}}\textbf{\textit{o}}\ (\textit{Informatividade}\ \text{é}\ \textit{Baixa})$$

R_2:

$$\textbf{\textit{Se}}\ (\textit{loc-len}\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_1\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_2\ \text{é}\ \textit{Médio})\ \textbf{\textit{Entã}}\textbf{\textit{o}}\ (\textit{Informatividade}\ \text{é}\ \textit{Média})$$

R_3:

$$\textbf{\textit{Se}}\ (\textit{loc-len}\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_1\ \text{é}\ \textit{Baixo})\ \textit{e}\ (K_2\ \text{é}\ \textit{Alto})\ \textbf{\textit{Entã}}\textbf{\textit{o}}\ (\textit{Informatividade}\ \text{é}\ \textit{Alta})$$

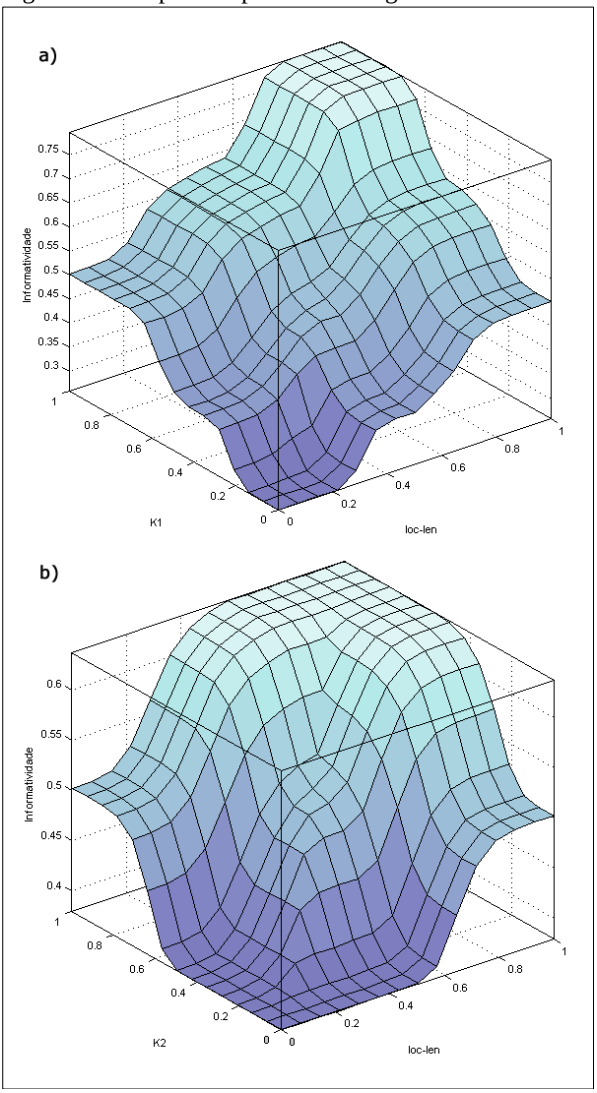
A base de regras usa variáveis linguísticas como seus antecedentes e consequentes. A regra três tem quatro variáveis linguísticas (*loc-len*, K_1, K_2 *e Informativiade*) que assumem os valores de *Baixo, Baixo, Alto e Baixa*, respectivamente. As três primeiras variáveis expressam uma inferência ou desigualdade que deve ser satisfeita, são os antecedentes da regra. A última variável é o consequente, a saída caso a desigualdade do antecedente na regra for satisfeita.

A formação lógica das regras é apresentada em um mapa de superfície na Figura 12. O mapa de superfície em 3D informa o quanto relevante é para uma saída *z* (*Informatividade*) a combinação de duas

entradas *x* (*loc-len*) e *y* (*K1* na Figura 12a, *K2* na Figura 12b), em um plano.

A base de regras linguísticas do FSumm encontra-se no Apêndice A.

Figura 12 – Mapa de superfície das regras



Fonte: Elaborada pelo autor

As funções de pertinência utilizadas para as variáveis linguísticas são do tipo gaussianas. Inicialmente procurou-se trabalhar com funções triangular/trapezoidal, mas no caso do FSumm esse tipo de função comprometeu o resultado final da análise *fuzzy*, pois algumas sentenças obtiveram o mesmo grau de informatividade. Valores repetidos de informatividade podem levar a uma indecisão no momento da seleção das sentenças para compor o sumário.

A função gaussiana é composta por três parâmetros: *a* (altura do pico da curva), *b* (a posição do centro do pico) e *c* (controla a largura da curva).

(25)

$$f(x) = a \times e^{-\frac{(x-b)^2}{2c^2}}$$

A Tabela 1 lista as funções de pertinências dos conjuntos com os parâmetros que assumem valores do universo de discurso entre 0 a 1. Os parâmetros foram ajustados por meio de experimentação manual e conforme a distribuição dos dados de uma amostra do corpus TeMário.

Tabela 1 – Descrição das variáveis e parâmetros das funções de pertinência

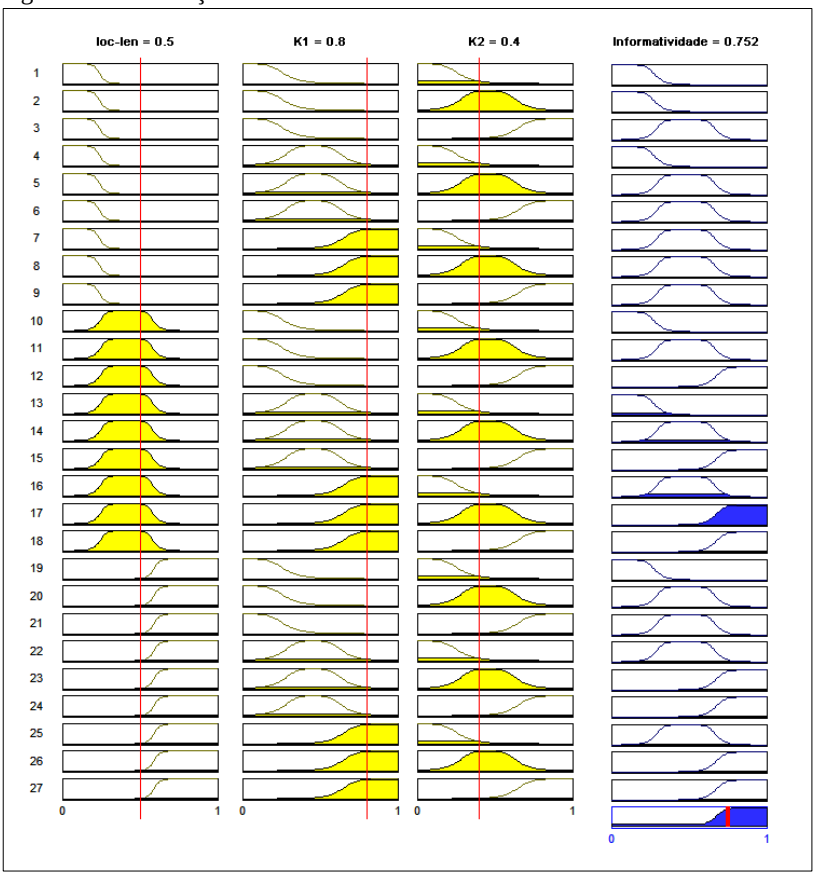
Análise fuzzy	Variável	Tipo	Variável linguística	Faixa de parâmetros da função [a; b; c]
Posição e comprimento	loc-len	entrada	Baixo	[0,241; 2,040; 0,046]
			Médio	[0,173; 2,810; 0,462]
			Alto	[0,387; 3,118; 1,030]
Keyword	K1	entrada	Baixo	[0,244; 2,410; 0,008]
			Médio	[0,159; 2,675; 0,411]
			Alto	[0,318; 3,465; 0,889]
Keyword	K2	entrada	Baixo	[0,258; 2,409; 0,008]
			Médio	[0,186; 1,983; 0,456]
			Alto	[0,318; 2,814; 0,952]
Classificação	Informatividade	saída	Baixa	[0,253; 4,160; 0,029]
			Média	[0,195; 3,117; 0,470]
			Alta	[0,305; 4,560; 0,975]

Fonte: Elaborada pelo autor

A Figura 13 simula o exemplo em que uma sentença recebe a pontuação $loc\text{-}len = 0{,}5$, $K1 = 0{,}8$ e $K2 = 0{,}4$. As variáveis de entrada

que correspondem às métricas extrativas inferem ao sistema os valores que ativam as regras na base. As linhas na Figura 13 representam as regras e as colunas mostram como cada variável (entradas e saída) ativam as regras. O sistema gera o valor desfuzzyficado de saída $Informatividade = 0{,}752;$ o que classifica a sentença com uma alta informatividade.

Figura 13 – Simulação do sistema



Fonte: Elaborada pelo autor

A base de regras permite fazer previsões da importância de uma sentença através do tratamento matemático da vagueza expressa nos termos linguísticos dos conjuntos *fuzzy*.

Para exemplificar as operações de conjuntos *fuzzy*, observa-se a regra 17. Os valores de entrada ($loc\text{-}len = 0{,}5$, $K1 = 0{,}8$ e $K2 = 0{,}4$.) são escalonados e quantizados para graus de pertinência as conjuntos *Baixo, Médio e Alto*. Usando a equação 25 e os parâmetros de cada conjunto *fuzzy* na Tabela 1, têm-se:

$$\mu_{Baixo}(loc\text{-}len) = 0{,}070$$
$$\mu_{Médio}(loc\text{-}len) = 0{,}999$$
$$\mu_{Alto}(loc\text{-}len) = 0{,}123$$

$$\mu_{Baixo}(K1) = 0{,}003$$
$$\mu_{Médio}(K1) = 0{,}008$$
$$\mu_{Alto}(K1) = 0{,}999$$

$$\mu_{Baixo}(K2) = 0{,}117$$
$$\mu_{Médio}(K2) = 0{,}991$$
$$\mu_{Alto}(K2) = 0{,}042$$

R_{17}: $\boldsymbol{Se}$ ($loc\text{-}len$ é $Médio$) e (K_1 é $alto$) e (K_2 é $Médio$)
$\boldsymbol{Então}$($Informatividade$ é $Alta$)

$$\mu_{Médio}(loc\text{-}len) \wedge \mu_{Alto}(K1) \wedge \mu_{Médio}(K2)\ ;$$
$$\Rightarrow Informatividade \text{ é } 0{,}991$$

Os antecedentes da regra são concatenados pelo conectivo *e*, que implica no operador *min*. Todas as saídas das regras são agregadas pelo método *max*, e gerado um número real pela equação 16.

Métodos tradicionais de sumarização extrativa lidam com o conhecimento de forma absoluta com base nas medidas das sentenças. No FSumm, as medidas são qualificadas por meio de análise *fuzzy* (quantificação dos conjuntos *fuzzy*, definição dos parâmetros das funções de pertinência, entre outros), para então estimar o grau de informatividade das sentenças, por meio de um valor desfuzzyficado.

4 AVALIAÇÃO E ANÁLISE DOS RESULTADOS

De maneira geral, a avaliação de um método de ST requer a análise do resultado do processo de sumarização, o sumário. Existem duas formas de avaliação de sumários: a automática e a humana. A avaliação automática compreende uma ferramenta que determina a informatividade do sumário. A humana determina aspectos da qualidade do sumário que não podem ser avaliados por uma ferramenta automática, como a coesão, gramaticalidade e coerência.

A fim de minimizar o trabalho e o tempo despendido na avaliação de sumários e permitir que os resultados obtidos fossem comparados aos de outros sistemas, para o presente trabalho, a qualidade dos sumários gerados foi avaliada em termos de informatividade.

A avaliação conduzida segue a classificação definida por Antiqueira (2007) como sendo intrínseca, pois os sumários são avaliados isoladamente; *black-box*, pois módulos internos não são considerados na avaliação, apenas a entrada e a saída é avaliada; *off-line*, pois é realizada de forma automática, sem julgamento humano; e comparativa, já que o resultado de outros sistemas são considerados.

A ferramenta ROUGE foi utilizada para avaliar a similaridade entre os sumários gerados pelos sistemas e os sumários de referência. As métricas ROUGE apresentam grande correlação com a avaliação humana e são baseadas na co-ocorrência de n-gramas (sequência de palavras e pares de palavras), e não de sentenças, entre o sumário automático e um ou mais sumário de referência (LIN, 2004).

Os sumários de referência não costumam ser do tipo extrativo, onde sentenças do texto-fonte são selecionadas para compor o modelo, mas apresentam os principais segmentos informativos que estão associados ao texto-fonte.

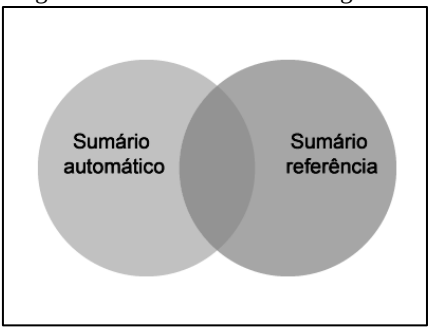
Em virtude do tamanho pequeno do corpus de estudo utilizado nesta dissertação, apenas a métrica ROUGE-1 foi selecionada para avaliar a informatividade dos sumários automáticos, em detrimento as demais.

ROUGE-1 é a divisão do número de unigramas (1-$grama$) que co-ocorrem no sumário automático ($Sum.aut.$) e nos sumários de referência ($Sum.ref.$), pelo número total do conjunto de unigramas presentes nos sumários de referência.

Portanto, a avaliação baseada em ROUGE consegue determinar se os mesmos conceitos gerais estão presentes em um sumário automático e em sumários de referência, mas não consegue determinar se os resultados encontrados são coerentes.

Os resultados de ROUGE produzem três medidas: precisão, cobertura e medida-f, sendo definidas da seguinte maneira:

Figura 14 – Co-ocorrência de n-gramas



Fonte: Elaborada pelo autor

$$Precisão = \frac{|n\text{-}gramas \in \{Sum.aut.\} \cap\ n\text{-}gramas\ \in \{Sum.ref.\}|}{|n\text{-}gramas \in Sum.aut.|} \quad (26)$$

$$Cobertura = \frac{|n\text{-}gramas \in \{Sum.aut.\} \cap\ n\text{-}gramas\ \in \{Sum.ref.\}|}{|n\text{-}gramas \in Sum.ref.|} \quad (27)$$

$$Medida\text{-}f = \frac{2 \times (Precisão\ \times Cobertura)}{Precisão + Cobertura} \quad (28)$$

A precisão expressa a proporção de n-gramas coincidentes entre o sumário automático e o de referência em relação ao número de n-gramas do sumário automático. Já a cobertura expressa a proporção de n-gramas coincidentes entre o sumário automático e o de referência em relação ao número de n-gramas do sumário de referência. Como a precisão e a cobertura são complementares, utiliza-se a medida-f que as agrupa em um único valor. Precisão e cobertura variam de 0 a 1, sendo que quando a precisão = 1 (100%), todas as n-gramas do sumário automático estão presentes no sumário de referência.

4.1 CORPUS

Para testar o método proposto, um corpus de domínio público com textos no idioma português foi selecionado. O corpus TeMário está

disponível pelo projeto Linguateca[13], que se destina ao processamento computacional da língua portuguesa. O corpus TeMário é composto por 100 textos jornalísticos, extraídos da Folha de São Paulo e Jornal do Brasil, igualmente distribuídos em cinco seções, totalizando 61.412 palavras e média de 29,37 sentenças. Os textos foram analisados por um especialista humano que assumiu o papel de sumarizador e leitor. O especialista produziu o um sumário de referência para cada texto-fonte, observando a restrição de 25% a 30% do tamanho do texto-fonte, e anotou quais as principais sentenças que indicavam a ideia principal (PARDO; RINO, 2003).

A escolha do corpus TeMário deu-se ao gênero jornalístico de seus textos, pois a linguagem é de fácil compreensão; a disponibilidade dos sumários de referência criados por um especialista em linguística e a possibilidade de uma avaliação comparativa, pois o corpus já foi utilizado na avaliação de outros sistemas de sumarização.

4.2 DEFINIÇÕES DOS EXPERIMENTOS

São três os experimentos que foram definidos nesta dissertação: análise de uma amostra do corpus, desempenho do FSumm com os textos do corpus e desempenho com um texto de natureza científica.

a) Análise de uma amostra do corpus: consistiu em analisar uma amostra de dados do corpus TeMário para validar a métrica proposta como um modelo à sumarização extrativa. Para este experimento criou-se um *script* que extrai dos textos-fonte a posição de cada sentença, a quantidade de termos, a média dos termos e o desvio padrão. Essas informações foram manipuladas em planilha eletrônica e posteriormente analisadas em softwares estatísticos – *Curve Fitting Tool* do MatLab e no R.

b) Desempenho do FSumm com os textos do corpus: o experimento baseou-se na avaliação com foco na informatividade dos sumários gerados pelos métodos *Baseline-0 (gold standard* para o TeMário*), Baseline-1, Baseline-2,* FSumm e GistSumm.

Segundo Christiansen (2014), o sumário *baseline* pode ser gerado de inúmeras formas, dependendo da natureza dos textos, do método de sumarização e da avaliação. Algumas das maneiras para a construção de sumários *baseline* são a seleção das primeiras sentenças do texto-fonte, a seleção aleatória de sentenças, ou ainda, a seleção de sentenças por meio da manipulação de uma característica textual.

[13] http://www.linguateca.pt/Repositorio/TeMario/

O método *Baseline-0* seleciona apenas as frases que aparecem no início do texto. A razão para a escolha dessa abordagem na geração de sumários *baseline* é pela natureza dos textos e pelo desempenho. De acordo com Nenkova e McKeown (2011), os textos jornalísticos apresentam as partes mais importantes no início do artigo e, por convenção, sumários *baseline* apresentam resultados que servem de referência na comparação com outros sistemas de sumarização.

O *Baseline-1* utiliza as métricas loc, len, K_1 e K_2 pela maneira tradicional, selecionando as sentenças de maior pontuação dada pela soma das métricas. O *Baseline-2* seleciona as sentenças da mesma forma que o método anterior, porém emprega a loc e len de forma correlacionada no modelo de regressão, além de K_1 e K_2.

O FSumm emprega as métricas loc-len, K_1 e K_2 inferidas pela lógica *fuzzy*. As sentenças selecionadas devem representar uma inforatividade Alta/Média. O GistSumm seleciona as sentenças com base na frequência das palavras, cuja métrica também é utilizado no FSumm, e destina-se a sumarização de texto em língua portuguesa (PARDO; RINO; NUNES, 2003).

O tamanho dos sumários produzidos pelos sistemas é proporcional ao número de palavras do texto-fonte, e não ao número de sentenças, sendo aplicada uma taxa de compressão de 30%.

Os sumários gerados pelos sistemas foram avaliados pela ROUGE. A ferramenta de avaliação ROUGE é desenvolvida em linguagem de programação Perl e para a sua execução é necessário instalar alguns módulos Perl. Após a instalação dos módulos para o sistema Windows, as configurações básicas de ROUGE foram utilizadas para comparar os sumários produzidos pelos sistemas com os sumários de referência do corpus, como demonstra o Quadro 6.

Quadro 6 – Configurações utilizadas na avaliação com ROUGE

```
-a: avalia todos os sistemas
-e: é o diretório onde os arquivos dos sumários são encontrados
-n: calcula as medidas ROUGE-N
-x: não calcular ROUGE-L
-c: especifica o intervalo de confiança (95%)
-r: especifica o número de pontos de amostragem
-f: seleciona a fórmula de escore (‘A’ => média do modelo)
-p: importância relativa da precisão e cobertura das medidas ROUGE

>ROUGE-1.5.5.pl -e data -n 2 -x -c 95 -r 1000 -f A -p 0.5 -a settings.xml
```

Fonte: Elaborado pelo autor

Antes de executar ROUGE, os sumários necessitam ser formatados e convertidos em arquivos do tipo html. A partir do *script prepare4rouge*[14] os sumários produzidos pelos sistemas e os considerados padrão-ouro foram transformados para o formato adequado e construída a estrutura necessária para os testes com a ferramenta de avaliação.

c) Desempenho do FSumm com um texto científico: o FSumm é aplicável a qualquer corpus, no entanto é necessário estimar os parâmetros da métrica *loc-len* quando os textos não forem do tipo artigos de jornal.

Apesar dos parâmetros da *loc-len* terem sido estimados utilizando textos jornalísticos, o FSumm também foi aplicado de maneira exploratória considerando uma parte do textual desta dissertação.

4.3 ANÁLISE DOS RESULTADOS

Os resultados obtidos nos experimentos são analisados e discutidos nas próximas seções. O primeiro experimento (Seção 6.3.1) demonstra que a convergência das métricas *loc* e *len* para uma única médica é comprovada por um modelo que apresenta um alto coeficiente de determinação. O segundo experimento (Seção 6.3.2) mostra que o desempenho do FSumm supera o desempenho de métodos tradicionais (*Baseline-1 e Baseline-2*). O terceiro experimento (Seção 6.3.3) apresenta o desempenho dos sistemas para um texto de natureza científica e aponta considerações sobre o processo de avaliação intrínseca baseado em n-gramas.

4.3.1 Especificação do modelo

Uma amostra de 79 textos do corpus TeMário foi selecionada aleatoriamente por sorteio, ao nível de confiança de 95%. A amostra foi submetida à etapa de pré-processamento e, após isso, definiu-se o conjunto de dados utilizados para compor o modelo de regressão multivariada.

Como variável resposta foi utilizada a medida de posição e comprimento *loc-len*, e como variáveis exploratórias foram utilizadas a quantidade de termos da sentença t_{siy} e o ranque de posição *loc*.

A análise dos dados revelou o seguinte modelo de regressão:

[14] http://kavita-ganesan.com/content/prepare4rouge-script-prepare-rouge-evaluation

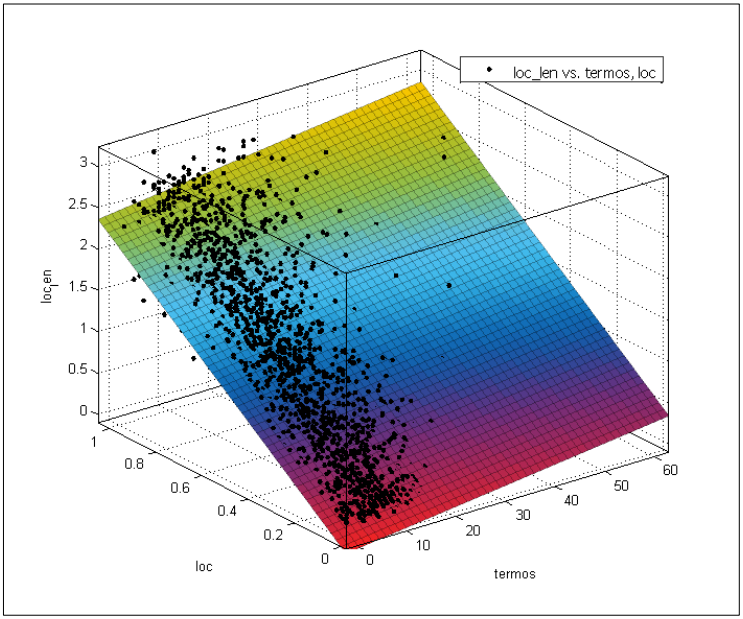
$$loc\text{-}len = -0{,}084 + 0{,}008\, t_{siy} + 2{,}344\, loc \quad (29)$$

O teste do modelo apresentou um coeficiente de determinação $R^2 = 0{,}954$, erro padrão de 0,15 e p-valor $< 0{,}01$. Sendo assim, rejeita-se a hipótese H_0: o modelo de regressão não é válido, logo $R^2 = 0$ e se afirma a hipótese H_1: $R^2 \neq 0$. Portanto, há uma forte associação entre a variável resposta e às variáveis exploratórias do modelo.

O incremento de 1 termo na sentença causa o incremento esperado de 0,008 na *loc-len* e o incremento de 1 na *loc* causa o incremento esperado de 2,344 na *loc-len*.

Através da análise dos coeficientes estimados para o modelo é possível verificar que a constante 0,084 influencia negativamente no resultado da *loc-len*. As variáveis t_{siy} e *loc* influenciam positivamente. A Figura 15 ilustra o mapa de superfície do modelo.

Figura 15 – Mapa de superfície do modelo



Fonte: Elaborada pelo autor

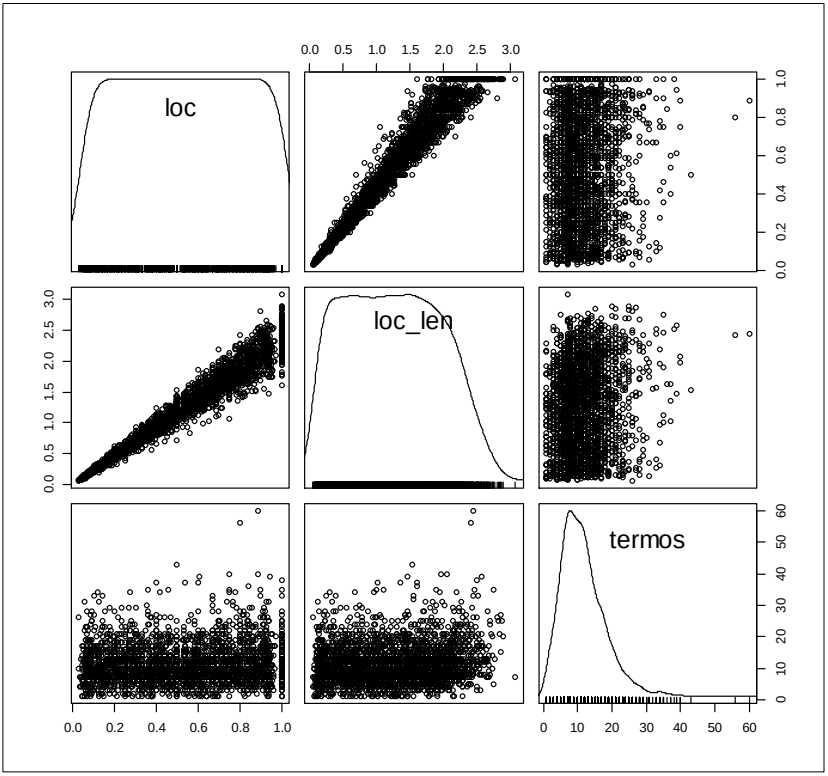
A cor no mapa de superfície indica a qualidade de aderência dos dados ao modelo. A cor vermelha corresponde ao menor desvio entre a superfície ajustada e os dados, enquanto que a cor amarela ao maior.

Também é possível notar na Figura 15 que há valores atípicos (*outliers*). São sentenças com o número de termos excessivamente superior as demais, mesmo tendo passado por uma etapa de pré-processamento.

No modelo estimado o parâmetro da variável *loc* é bastante alto, pois refere-se a uma característica textual de grande importância para textos jornalísticos (NENKOVA; MCKEOWN, 2011).

Embora o modelo proposto tenha apresentado valores de teste estatisticamente significativos, ainda pode ser aprimorado. A Figura 16 apresenta uma matriz com as variáveis loc, t_{siy} e $loc\text{-}len$. Na diagonal principal encontram-se os gráficos da distribuição da frequência das variáveis e nas demais linhas e colunas, os gráficos de dispersão.

Figura 16 – Matriz de dispersão



Fonte: Elaborada pelo autor

A dispersão dos dados aumenta conforme o valor da loc também aumenta (gráfico a_{21}), portanto as sentenças com valores de loc próximos a 1 estão mais sujeitas ao erro.

A conversão das curvas de distribuição das frequências da loc (gráfico a_{11}) e da quantidade de termos (gráfico a_{33}) pode ser observada no gráfico da $loc\text{-}len$ (gráfico a_{22}). O gráfico da $loc\text{-}len$ apresenta-se mais semelhante à curva de distribuição da frequência da loc. Possivelmente, se a quantidade de termos possuir maior influência no modelo, maior será a influencia dessa variável na curva de distribuição da frequência da $loc\text{-}len$.

4.3.2 Desempenho do FSumm com os textos jornalísticos

Os resultados obtidos a partir da avaliação com a ferramenta ROUGE são dados em termos de precisão (*P* – indica o quão próximo o sumário automático está do sumário de referência), cobertura (*C* – indica o quanto de informação do sumário de referência está no sumário automático) e medida-f (f – medida que une a *P* e *C*, portanto, uma medida de eficiência).

Com essas medidas, procurou-se avaliar o grau de informatividade dos sumários automáticos, em relação aos sumários de referência. A ideia de avaliar a informatividade é verificar se o conteúdo de um bom sumário também está presente no sumário automático.

Neste caso, a coesão, coerência, gramaticabilidade ou qualquer outra propriedade textual dos sumários automáticos diferente da informatividade, não são consideradas na avaliação com ROUGE. A Tabela 2 ilustra os resultados da co-ocorrência de unigramas entre os sumários produzidos pelos cinco sistemas e os sumários de referência do corpus TeMário.

Tabela 2 – Resultados da ROUGE-1

SISTEMAS	ROUGE-1 (unigrama)			
	Precisão (P)	*Cobertura (C)*	*Medida-f*	*IC para a Medida-f (95%)*
*Baseline-0**	0,477	0,491	0,479	0,467 – 0,491
Baseline-1	0,458	0,477	0,463	0,450 – 0,475
Baseline-2	0,457	0,480	0,464	0,452 – 0,477
GistSumm	0,496	0,429	0,456	0,445 – 0,468
FSumm	0,462	0,475	0,465	0,453 – 0,480

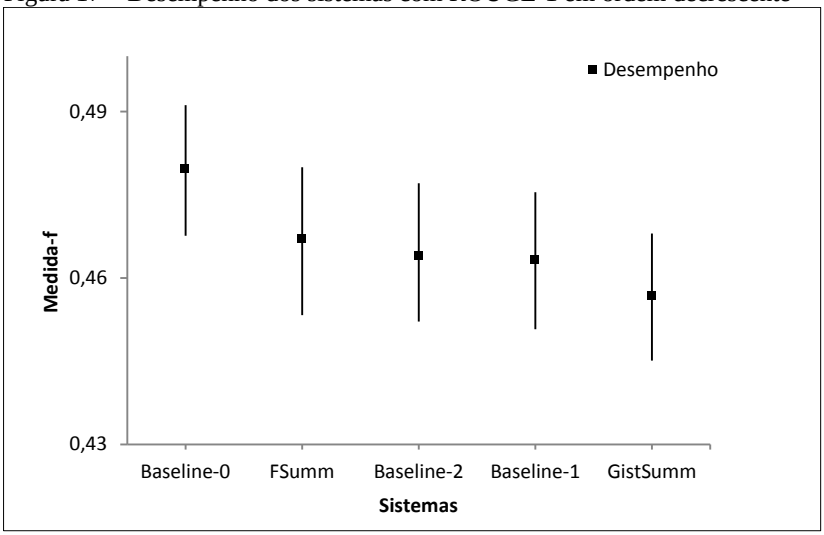
*Padrão ouro para o TeMário
Fonte: Elaborada pelo autor

Nos resultados da avaliação com ROUGE-1 observa-se que o GistSumm é o sistema que apresentou a maior precisão, com P = 0,496, mas também a menor cobertura C = 0,429. O GistSumm é um sumarizador que primeiro escolhe a sentença considerada a mais importante do texto-fonte, para então selecionar as sentenças que possuam alguma semelhança a ela. O método de sumarização presente no GistSumm mostrou-se preciso, porém, com a menor cobertura.

O *Baseline-0* demostra-se mais informativo que os demais sistemas. Com uma cobertura C = 0,491, o *Baseline-0* consegue selecionar mais n-gramas dos sumários gerados que ocorrem no sumário de referência.

Embora o desempenho entre os sistemas seja próximo, os resultados referentes à aplicação da métrica ROUGE-1 demonstram que o sistema com melhor desempenho foi o *Baseline-0*, com f = 0,479. O FSumm não conseguiu bater o valor de desempenho do sistema *Baseline-0*, mas foi o segundo melhor sumarizador, com f = 0,465 seguido pelo *Baseline-2*, *Baseline-1* e GistSumm, com f = 0,464, f = 0,463 e f = 0,456, nesta ordem. A Figura 17 ilustra a classificação dos sistemas conforme o desempenho (f) e os intervalos de confiança (*IC*) com ROUGE-1.

Figura 17 – Desempenho dos sistemas com ROUGE-1 em ordem decrescente



Fonte: Elaborada pelo autor

O *IC* e o desempenho dos sistemas mostram que os sistemas são semelhantes, sendo o FSumm o mais próximo ao *Basiline-0* e o GistSumm o mais distante.

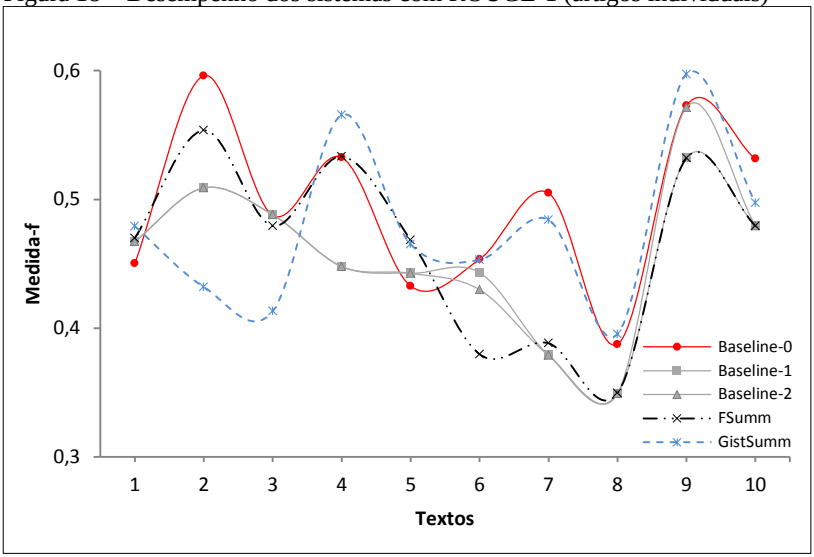
O sistema *Baseline-0* é considerado simples, e serve de base na avaliação para o sistema proposto. De maneira geral, o melhor desempenho do *Baseline-0* em relação aos demais sistemas não é surpresa, pois como discutido na Seção 6.2, os textos do corpus TeMário são de natureza jornalística e as principais informações estão no início do texto.

Ao analisar o *IC* e o desempenho dos sistemas que implementam as métricas propostas nesta dissertação, o FSumm apresenta-se melhor quando comparado ao *Baseline-1* e *Baseline-2*. O *Baseline-2*, que implementa as métricas correlacionadas, obteve um desempenho pouco superior ao *Baseline-1*, onde as métricas são usadas sem a correlação. Portanto, ainda que a diferença no desempenho dos sistemas seja sutil, a correlação das métricas apresenta-se como uma forma de lidar com a quantidade de caraterísticas textuais sem comprometer o desempenho do sistema.

Um sistema que seleciona as primeiras sentenças de um artigo de jornal cria uma base forte e difícil de superar. No entanto, quando analisado o desempenho dos sistemas em artigos individuais da coleção

do corpus, vê-se que em alguns casos o FSumm, o GistSumm e o *Baseline-2* foram melhores que o *Baseline-0*. A Figura 18 mostra o desempenho dos sistemas com ROUGE-1 para uma amostra de dez artigos individuais.

Figura 18 – Desempenho dos sistemas com ROUGE-1 (artigos individuais)



Fonte: Elaborada pelo autor

A maior variação no desempenho foi do artigo 2, 16% em relação ao valor mais alto e mais baixo. O artigo 1 apresentou a menor variação, 3%. O texto-fonte do artigo 1 e os sumários automáticos e de referência encontram-se no Apêndice B.

4.3.3 Desempenho do FSumm com um texto científico

Para entender melhor os resultados quantitativos obtidos com a utilização da ferramenta ROUGE, buscou-se analisar o desempenho do FSumm explorando um texto cientifico. O texto-fonte pertence à Seção 1.2 desta dissertação e o sumário de referência provêm do Resumo. O Quadro 7 ilustra o sumário de referência e os produzidos pelos métodos *Baseline-0* e FSumm.

Quadro 7 – Sumários produzidos pelos métodos a partir de um texto científico

Referência: *A sumarização automática de texto procura condensar o conteúdo do documento, extraindo as informações mais relevantes. Esse processo normalmente é executado através de métodos computacionais, como o método estatístico e o linguístico. O rápido desenvolvimento das tecnologias emergentes e a crescente quantidade de informação disponível inserem novos desafios para esta área de pesquisa.* ***Um desses desafios está na identificação das sentenças mais informativas no momento da geração do sumário****. Como a tarefa de sumarizar informações de texto trás consigo a incerteza inerente à linguagem natural, a Lógica fuzzy pode ser aplicada nessa tarefa para contribuir nos resultados gerados.*
Baseline-0: *[**s1**] Grande parte da informação está na forma textual e devido ao volume identificar o que é relevante e de forma coerente remetem a capacidade de sumarizar. [**s2**] O fluxo de informações em um documento não é uniforme o que significa que algumas partes são mais importantes do que outras. [**s3**]* ***O grande desafio na sumarização encontra-se em distinguir quais são as partes mais informativas****. [**s4**] Quais são as propriedades textuais que devem ser retidas no sumário? [**s5**] Quais as melhores métricas para alcançar o conteúdo de documentos? [**s6**] Essas perguntas são subjetivas e difíceis de responder, pois a pontuação das características trás consigo a incerteza e a imprecisão.*
FSumm: *[**s10**] A lógica fuzzy aplicada a ST por meio de funções de pertinência é capaz de lidar com as diferentes características do texto consideradas com igual importância nos métodos de sumarização clássicos. [**s11**] Com o intuito de explorar um formalismo para representar a incerteza por imprecisão na extração de informação a principal contribuição desta pesquisa é um novo método de sumarização fuzzy. [**s12**] A hipótese levantada é a de uma vez que são utilizadas métricas para identificar o conteúdo informativo de textos algumas métricas são mais importantes e a lógica fuzzy é capaz de modelar tal problema para a construção de um sumário.*

Fonte: Elaborado pelo autor

No Quadro 7 a sentença destacada "*O grande desafio na sumarização encontra-se em distinguir quais são as partes mais informativas*" no sumário automático, transmite a mesma ideia que a sentença do sumário de referência, "*Um desses desafios está na identificação das sentenças mais informativas no momento da geração do sumário*". As palavras descritivas das sentenças podem ser representadas da seguinte forma:

$$s1\colon [desafio, \boldsymbol{sumariza}\text{ção}^{\mathbf{1}}, encontra, \boldsymbol{distinguir}^{\mathbf{2}}, \boldsymbol{parte}^{\mathbf{3}}, informativa]$$

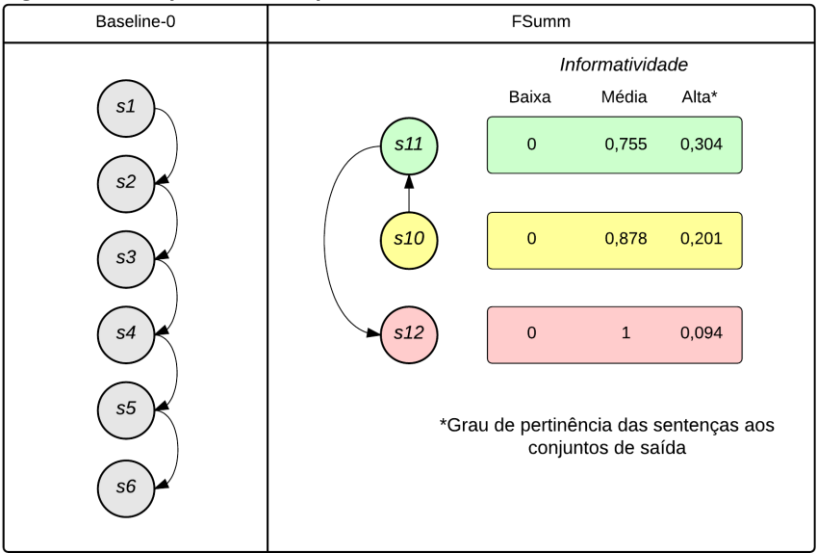
$$s2\colon \begin{bmatrix} desafio, \boldsymbol{identifica}\text{ção}^{\mathbf{2}}, \boldsymbol{sentença}^{\mathbf{3}}, informativa, momento, \\ \boldsymbol{gera}\text{ção}, \boldsymbol{sum}\text{ário}^{\mathbf{1}} \end{bmatrix}$$

Nas sentenças $s1$ e $s2$ é possível observar que as palavras de índices iguais apresentam o mesmo significado nas respectivas frases, porém não poderiam ser considerados na avaliação por ROUGE, pois são unigramas diferentes. Desse modo, é preciso investigar as possibilidades das relações semânticas dos termos no processamento da

avaliação automática de textos em língua portuguesa. Se a ferramenta ROUGE considerasse um dicionário de sinônimos para textos em língua portuguesa, os valores de desempenho da tabela 2 poderiam ser melhores.

A Figura 19 ilustra a construção dos sumários automáticos do quadro anterior. O *Baseline-0* seleciona as sentenças conforme a ordem do posicionamento no texto-fonte. O FSumm seleciona as sentenças pelo ranqueamento obtido com a desfuzzyficação.

Figura 19 – Seleção das sentenças no *Baseline-0* e FSumm



Fonte: Elaborada pelo autor

No método FSumm a sentença $s11$ aparece no topo do ranque com um valor desfuzzyficado igual a 0,641. O valor especificado representa o grau de informatividade da $s11$, e que corresponde aos graus de pertinência $\mu_{Alta} = 0{,}304$ ao subconjunto Alta; $\mu_{Média} = 0{,}755$ ao Média e grau zero ao Baixa. Os termos Baixa, Média e Alta referem-se a classificações linguísticas da variável de saída Informatividade.

A diminuição da rigidez dos limites dos subconjuntos permite lidar com a pontuação das sentenças de maneira incerta. Assim, a $s11$ apresenta a possibilidade de pertencer mais ao subconjunto Média do que ao Alta, porém no momento de classificar a informatividade da sentença, ambos os subconjuntos são considerados.

Assim, os resultados mostram que é possível tratar a informatividade das sentenças por uma escala de valores que varia do menor grau (0) até o maior (1), e que pode ser utilizada como uma medida para a seleção das sentenças.

5 CONSIDERAÇÕES FINAIS

Nesta dissertação procurou-se alinhar a sumarização de texto extrativa e a lógica *fuzzy* para desenvolver um método automático que produzisse sumários, chamado de FSumm.

O método FSumm é composto por três processos-chaves. O primeiro processo trabalha com as etapas de pré-processamento de texto; o segundo, com as métricas extrativas do texto; e o terceiro, com a transformação das métricas extrativas em uma medida *fuzzy*. A medida *fuzzy* atribui uma relevância as informações de um texto que são utilizadas para a construção do sumário.

As principais características textuais, com base na literatura, definiram as métricas extrativas. Duas das métricas foram combinadas para compor uma nova métrica. A nova métrica foi estimada em um modelo de regressão multivariada e validada pelos resultados dos experimentos.

A avaliação do método proposto deu-se por uma ferramenta automatizada que analisa a informatividade dos sumários automáticos em relação aos sumários produzidos por humanos. Um conjunto de artigos de jornais de língua portuguesa formou o corpus de textos aplicado nos experimentos.

Dentre os métodos sumarizadores utilizados nos experimentos, o *Baseline-0* apresentou o melhor desempenho geral, comprovando que para artigos jornalísticos, a seleção das primeiras sentenças do texto-fonte forma uma base difícil de superar. Ainda assim, o FSumm foi o método que mais próximo chegou do desempenho geral do *Baseline-0*. Quando foram analisados os artigos individuais, em alguns momentos o desempenho do *Baseline-0* é inferior aos dos outros métodos, porém esse aspecto precisa ser melhor investigado.

Por utilizar a lógica *fuzzy*, o FSumm permite que a relevância de certas partes do texto que constituem um sumário sejam consideradas de forma imprecisa, pois são qualificadas em subconjuntos representados por termos linguísticos. Além disso, a decisão de selecionar ou não uma sentença considera simultaneamente a possibilidade de pertencer a mais de uma categoria dos subconjuntos.

O método proposto nesta dissertação demonstra que as métricas extrativas que derivam das características do texto podem ser aperfeiçoadas e incorporadas ao processo de sumarização gerando resultados melhores do que os de métodos tradicionais, como os implementados no *Baseline-1* e *2*.

Embora a correlação de características textuais seja mais complexa do que a forma tradicional de sumarização, traz como vantagens a simplificação do problema, pois independente da abordagem adotada pelo método de sumarização, diminui-se a quantidade de variáveis; e especificamente em sistemas com lógica *fuzzy*, impacta na quantidade de regras de associação, consequentemente no desempenho do sistema de inferência.

O objetivo desta dissertação que é propor um método fuzzy para a sumarização de texto por meio de métricas extrativas foi alcançado, assim como os objetivos específicos relacionados à identificação das características do texto, a definição e aperfeiçoamento de métricas e a avaliação do método.

Em termos de limitações, o método proposto desconsidera qualquer tipo de análise semântica ou morfossintática mais profunda. Ao introduzir-se uma tarefa que envolva a morfossintática, como por exemplo, a lematização na fase de pré-processamento, os resultados do desempenho dos sistemas podem ser diferentes.

A sumarização do método FSumm é monodocumento, e na avaliação buscou-se verificar a informatividade dos sumários. Assim, a utilização do FSumm para a sumarização multidocumento fica condicionada a avaliação de outras propriedades do texto (como a coerência e a redundância) e a implementação de uma métrica de ordenação das sentenças.

Os experimentos conduzidos nesta dissertação levaram em consideração apenas um corpus com textos jornalísticos em língua portuguesa. Como as métricas e os processos de pré-processamento empregados no método proposto independem do idioma dos textos, acredita-se que o desempenho do FSumm, para um corpus em língua inglesa, não seja comprometido. Ainda assim, faz-se necessário investigar o desempenho do FSumm com a inclusão de outras características textuais, assim como para corpos com mais textos, o que permitiria a utilização de uma avaliação também por bigrama.

A correlação das características posição e comprimento comprovaram-se por meio da regressão linear. Quando a técnica de regressão linear é aplicada, os parâmetros estimados do modelo resultante são limitados aos dados. Os dados que serviram para estabelecer a correlação entre as características foram extraídos do corpus TeMário, portanto deve-se investigar a correlação da posição e o comprimento das sentenças, assim como outras características, em coleções de texto de diferente idioma e natureza (texto acadêmico, por exemplo).

A base de regras pode ser aperfeiçoada conforme surjam novos experimentos ou diferentes situações a serem modeladas. O uso de algoritmos genéticos ou aprendizagem de máquina poderiam implementar funções de pertinência e quantificar conjuntos dinamicamente, principalmente em situações em que os conhecimentos são relativamente vagos.

REFERÊNCIAS

AGGARWAL, C.; ZHAI, C. **Mining text data**. Springer, US, v. 4, p. 889–903, 2012.

ALGULIEV, R. et al. MCMR: Maximum coverage and minimum redundant text summarization model. **Expert Systems with Applications**, v. 38, n. 12, p. 14514–14522, 2011.

ALGULIEV, R. M. O.; ALYGULIEV, R. M. O. Automatic Text Documents Summarization through Sentences Clustering. **Journal of Automation and Information Sciences**, v. 40, n. 9, p. 53–63, 2008.

AMIGÓ, E.; GONZALO, J.; PEÑAS, A.; VERDEJO, F. QARLA: a framework for the evaluation of text summarization systems. In: **Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics**, p. 280–289, 2005.

ANTIQUEIRA, Lucas. **Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos**. 2007. Dissertação (Mestrado em Ciências Matemáticas e de Computação). Universidade de São Paulo, 2007.

ARIES, Abdelkrime; OUFAIDA, Houda; NOUALI, Omar. Using clustering and a modified classification algorithm for automatic text summarization. In: **IS&T/SPIE Electronic Imaging**. International Society for Optics and Photonics, v. 8658. p. 865811-9, 2013.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**: The Concepts and Technology behind Search. 2. ed. Addison-Wesley, Harlow, 2011.

BALABANTARAY, R. C. et al. Odia Text Summarization using Stemmer. **International Journal of Applied Information Systems (IJAIS)**, v. 1, n. 3, p. 21–24, 2012.

BALAGE FILHO, Pedro Paulo; PARDO, Thiago Alexandre Salgueiro; NUNES, Maria das Graças Volpe. Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm. **Série de Relatórios do NILC**. NILC-TR-07-11. 2007.

BARZILAY, R.; MCKEOWN, K. K. R. Sentence fusion for multidocument news summarization. **Computational Linguistics**, v. 31, n. September 2003, p. 34, 2005.

BATCHA, N. K.; ZAKI, A. M. Algebraic reduction in automatic text summarization–the state of the art. In: **International Conference on Computer and Communication Engineering (ICCCE)**, Kuala Lumpur, p. 1-6, 2010.

BAXENDALE, P. B. Machine-made index for technical literature - an experiment. **IBM Journal of Research and Development**, v. 2, p. 354–365, 1958.

BINWAHLAN, Mohammed Salem; SALIM, Naomie; SUANMALI, Ladda. Fuzzy swarm diversity hybrid model for text summarization. **Information processing & management**, v. 46, n. 5, p. 571-588, 2010.

BRANDOW, R.; MITZE, K.; RAU, L. Automatic condensation of electronic publications by sentence selection. **Information Processing & Management**, v. 31, n. 5, p. 675–685, 1995.

CAMARGO, Renata Tironi. **Investigação de estratégias de sumarização humana multidocumento**. 2013. 133f. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos.

CASTRO JORGE, Maria Lucía del Rosario; PARDO, Thiago Alexandre Salgueiro. Experiments with CST-based multidocument summarization. In: **Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing**. Association for Computational Linguistics. p. 74-82, 2010.

CHANDRA, Munehs; GUPTA, Vikrant; PAUL, Santosh Kr. A statistical approach for automatic text summarization by extraction. In: **Communication Systems and Network Technologies (CSNT)**, 2011 International Conference on. IEEE, p. 268-271, 2011.

CHRISTIANSEN, Karin. **Summarization Of Icelandic Texts**. 2014. 74 f. Projeto Dissertação (Mestrado) - Curso de Master of Science in Computer Science, School Of Computer Science, Universidade de Reiquiavique, Reiquiavique.

DAS, D.; MARTINS, A. F. T. A Survey on Automatic Text Summarization Single-Document Summarization**. Literature Survey for the Language and Statistics II course at CMU**, v. 4, p. 1–31, 2007.

DAS, Samarjit. Pattern Recognition using the Fuzzy c-means Technique. **International Journal of Energy**, **Information and Communications**, v. 4, n. 1, p. 1–14, 2013.

DOKO, A.; ŠTULA, M.; ŠERIĆ, L. Improved sentence retrieval using local context and sentence length. **Information Processing & Management**, v. 49, n. 6, p. 1301–1312, 2013.

EDMUNDSON, H. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264-285, 1969.

GIANNAKOPOULOS, G. et al. Summarization system evaluation revisited. **ACM Transactions on Speech and Language Processing**, v. 5, n. 3, p. 1–39, 2008.

GOULARTE, F.; WILGES, B.; NASSAR, S. M. Métricas de sumarização automática de texto em tarefas de um Ambiente Virtual de Aprendizagem. In: Anais do Simpósio Brasileiro de Informática na Educação. 2014. p. 752-761.

GUPTA, S.; NENKOVA, A.; JURAFSKY, D. Measuring importance and query relevance in topic-focused multi-document summarization. In: **Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions**. p. 193–196, 2007.

GUPTA, Vishal; LEHAL, Gurpreet Singh. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HANNAH, M. Esther; GEETHA, T. V.; MUKHERJEE, Saswati. Automatic extractive text summarization based on fuzzy logic: a sentence oriented approach. In: **Swarm, Evolutionary, and Memetic Computing**. Springer, p. 530-538, 2011.

HILBERT, Martin; LÓPEZ, Priscila. The world's technological capacity to store, communicate, and compute information. **Science** (New York, N.Y.), v. 332, n. 6025, p. 60-65, abr. 2011.

IBRAHIM, Ahmad. **Fuzzy logic for embedded systems applications**. Newnes, 2004.

JONES, K. Automatic summarizing: factors and directions. **Advances in automatic text summarization**, p. 1–21, 1999.

JONES, K. S. Automatic summarising: The state of the art. **Information Processing & Management**, v. 43, n. 6, p. 1449–1481, nov. 2007.

KATRAGADDA, Rahul. GEMS: generative modeling for evaluation of summaries. In: **Computational Linguistics and Intelligent Text Processing**. Springer, p. 724-735, 2010.

KIABOD, M.; DEHKORDI, M. N.; SHARAFI, M. A Novel Method of Significant Words Identification in Text Summarization. **Journal of Emerging Technologies in Web Intelligence**, v. 4, n. 3, p. 252–259, 2012.

KIANI, A.; AKBARZADEH, M. R. Automatic Text Summarization Using Hybrid Fuzzy GA-GP. In: **2006 IEEE International Conference on Fuzzy Systems**, p. 977–983, 2006.

KITCHENHAM, B. **Procedures for performing systematic reviews**. Keele, UK, Keele University, 2004.

KLIR, George; YUAN, Bo. **Fuzzy sets and fuzzy logic**. New Jersey: Prentice Hall, 1995.

KUPIEC, Julian; PEDERSEN, Jan; CHEN, Francine. A trainable document summarizer. In: **Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval**. ACM, p. 68-73, 1995.

KYOOMARSI, F.; KHOSRAVI, H.; ESLAMI, E.; DAVOUDI, M. Extraction-based text summarization using fuzzy analysis. **Iranian Journal of Fuzzy Systems**, v. 7, n. 3, p. 15–32, 2010.

KYOOMARSI, F.; KHOSRAVI, H.; ESLAMI, E.; DEHKORDY, P. K.; TAJODDIN, Asghar. Optimizing Text Summarization Based on Fuzzy Logic. **7th IEEE/ACIS International Conference on Computer and Information Science (ICIS '08)**, p. 347–352, 2008.

LEITE, Daniel; RINO, Lucia H. M. A Genetic Fuzzy Automatic Text Summarizer. In: **Anais do CSBC 2009**. p. 779–788, 2009.

LI, Hang; YAMANISHI, Kenji. Document classification using a finite mixture model. In: **Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, p. 39-47, 1997.

LIMA, J.; PARDO, T. Ordenação de Sentenças em Sumários Multidocumento. São Paulo. **Série de Relatórios do NILC**. NILC-TR-12-02. 2012. Disponível em: <http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_382.pdf>. Acesso em: 27 out. 2014.

LIN, Chin-Yew. Rouge: A package for automatic evaluation of summaries. In: **Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004**, Barcelona, Spain. p. 74-81, 2004.

LIN, C.; HOVY, E. The automated acquisition of topic signatures for text summarization. **In: Proceedings of the 18th conference on Computational Linguistics**. v. 1, p. 495–501, 2000.

LLORET, E.; PALOMAR, M. Text summarisation in progress: a literature review. **Artificial Intelligence Review**, v. 37, n. 1, p. 1–41, 2012.

LUHN, H. The automatic creation of literature abstracts. **IBM Journal of research and development**, p. 159–165, 1958.

LUO, W. et al. Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. **Knowledge-Based Systems**, v. 46, p. 33–42, 2013.

MANI, I. Summarization evaluation: An overview. In: **Proceedings of the NAACL 2001 Workshop on Automatic Summarization**. 2001.

MEGALA, S. S.; KAVITHA, A.; MARIMUTHU, A. Enriching Text Summarization using Fuzzy Logic. **International Journal of Computer Science & Information Technologies**, v. 5, n. 1, p. 863–867, 2014.

MEI, J.-P.; CHEN, L. SumCR: A new subtopic-based extractive approach for text summarization. **Knowledge and Information Systems**, v. 31, n. 3, p. 527–545, 2011.

MENDEL, Jerry M. Type-2 fuzzy sets and systems: an overview. **Computational Intelligence Magazine**, IEEE, v. 2, n. 1, p. 20-29, 2007.

MÓRO, R.; BIELIKOV, M. Personalized text summarization based on important terms identification. In: **Database and Expert Systems Applications (DEXA)**, 23rd International Workshop, IEEE, p. 131–135, 2012.

NENKOVA, A.; MCKEOWN, K. **Automatic Summarization. Foundations and Trends® in Information Retrieval**, v. 5, n. 3, p. 235–422, 2011.

NENKOVA, A.; PASSONNEAU, R. Evaluating content selection in summarization: The pyramid method. In: **HLT/NAACL**, 2004.

NENKOVA, Ani; PASSONNEAU, Rebecca; MCKEOWN, Kathleen. The pyramid method: Incorporating human content selection variation in summarization evaluation. **ACM Transactions on Speech and Language Processing (TSLP)**, v. 4, n. 2, p. 4, 2007.

NENKOVA, Ani; VANDERWENDE, Lucy. The impact of frequency on summarization. **Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101**, 2005.

PAICE, Chris D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In**: Proceedings of the 3rd annual ACM conference on Research and**

development in information retrieval. Butterworth & Co, p. 172-191, 1980.

PARDO, Thiago Alexandre Salgueiro; RINO, Lucia Helena Machado. TeMário: Um Corpus para Sumarização Automática de Textos. São Carlos. **Série de Relatórios do NILC**. NILC-TR-03-09. 2003.

PARDO, Thiago Alexandre Salgueiro; RINO, Lucia Helena Machado; NUNES, Maria das Graças Volpe. GistSumm: A summarization tool based on a new extractive method. In: **Computational Processing of the Portuguese Language**. Springer, p. 210-218, 2003.

PASSONNEAU, Rebecca J. Formal and functional assessment of the pyramid method for summary content evaluation. **Natural Language Engineering**, v. 16, n. 02, p. 107, 2009.

RADEV, Dragomir R.; HOVY, Eduard; MCKEOWN, Kathleen. Introduction to the special issue on summarization. **Computational linguistics**, v. 28, n. 4, p. 399-408, 2002.

RADEV, D.; TAM, D.; ERKAN, G. Single-document and multi-document summary evaluation using Relative Utility. In: **Proceedings of the ACM CIKM**, New Orleans, LA, p. 1–28, 2003.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. Graph-based methods for multi-document summarization: exploring relationship maps, complex networks and discourse information. In: **Computational Processing of the Portuguese Language**. Springer, p. 260-271, 2012.

RINO, Lúcia Helena Machado; PARDO, Thiago Alexandre Salgueiro. A Sumarização Automática de textos: principais características e metodologias. In: **Anais do XXIII Congresso da Sociedade Brasileira de Computação.** p. 203-245, 2003.

ROSS, T. **Fuzzy logic with engineering applications**. 3rd. ed. John Wiley & Sons: UK, 2010.

SCHMIDT, Stefan; STEELE, Robert; DILLON, Tharam S. Towards usage policies for fuzzy inference methodologies for trust and QoS

assessment. **Computational Intelligence, Theory and Applications**. Springer, p. 263-274, 2006.

SUANMALI, L.; BINWAHLAN, M. S.; SALIM, N. Sentence Features Fusion for Text Summarization Using Fuzzy Logic. In: **2009 Ninth International Conference on Hybrid Intelligent Systems**, v. 1, p. 142–146, 2009.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M. S. Fuzzy Logic Based Method for Improving Text Summarization. **International Journal of Computer Science and Information Security (IJCSIS)**. v. 2, n. 1, 2009.

SUANMALI, Ladda; SALIM, Naomie; BINWAHLAN, Mohammed Salem. Fuzzy genetic semantic based text summarization. In: **Ninth International Conference on Dependable, Autonomic, and Secure Computing**. Sydney, NSW, p. 1184-1191, 2011.

SUNEETHA, S. Automatic Text Summarization: the current state of the art. **International Journal of Science and Advanced Technology**, p. 283-293, 2011.

SVORE, Krysta Marie; VANDERWENDE, Lucy; BURGES, Christopher JC. Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. In: **Proceedings of the EMNLP-CoNLL**, p. 448–457, 2007.

TRATZ, S.; HOVY, E. Summarization evaluation using transformed basic elements. In: **Proceedings of the 1st Text Analysis Conference**, 2008.

VADLAPUDI, Ravikiran; KATRAGADDA, Rahul. Quantitative evaluation of grammaticality of summaries. **Computational Linguistics and Intelligent Text Processing**. Springer , p. 736-747, 2010.

WITTE, R.; BERGLER, S. Fuzzy clustering for topic analysis and summarization of document collections. **Advances in Artificial Intelligence**, n. 1, p. 1–12, 2007.

YANG, G. et al. The effectiveness of automatic text summarization in mobile learning contexts. **Computers & Education**, v. 68, p. 233–243, out. 2013.

YIH, W. et al. Multi-Document Summarization by Maximizing Informative Content-Words. In: **Proceedings of the International Joint Conference on Artificial Intelligence**, p. 1776–1782, 2007.

ZADEH, L. Fuzzy sets as a basis for a theory of possibility. **Fuzzy sets and systems**, v. 1, p. 3–28, 1999.

APENDICE A – Base de regras do Fsumm

Se	*loc-len*	**e**	*K1*	**e**	*K2*	**então**	*Informatividade*
R_1	baixo		baixo		baixo		baixa
R_2	baixo		baixo		médio		baixa
R_3	baixo		baixo		alto		baixa
R_4	baixo		médio		baixo		baixa
R_5	baixo		médio		médio		média
R_6	baixo		médio		alto		média
R_7	baixo		alto		baixo		média
R_8	baixo		alto		médio		média
R_9	baixo		alto		alto		média
R_{10}	médio		baixo		baixo		baixa
R_{11}	médio		baixo		médio		média
R_{12}	médio		baixo		alto		média
R_{13}	médio		médio		baixo		baixa
R_{14}	médio		médio		médio		média
R_{15}	médio		médio		alto		alta
R_{16}	médio		alto		baixo		média
R_{17}	médio		alto		médio		alta
R_{18}	médio		alto		alto		alta
R_{19}	alto		baixo		baixo		média
R_{20}	alto		baixo		médio		média
R_{21}	alto		baixo		alto		alta
R_{22}	alto		médio		baixo		média
R_{23}	alto		médio		médio		alta
R_{24}	alto		médio		alto		alta
R_{25}	alto		alto		baixo		alta
R_{26}	alto		alto		médio		alta
R_{27}	alto		alto		alto		alta

APENDICE B – Artigo 1

Como direcionar sua empresa para o cliente

É preciso inverter as estruturas para diminuir a distância entre o cliente e os que detêm o poder de decisão.

Empresa fadada ao insucesso tem duas caras: uma real, outra para o cliente. Na hora de vender, promessas; quando o cliente confere, decepções. É difícil encontrar o responsável, quando a empresa não é direcionada à satisfação total dos clientes.

Nas estruturas tradicionais de empresas, onde o mando predomina sobre a responsabilidade individual, não é possível sequer aprender em cima dos próprios erros. Faz-se de tudo para que não haja registro do erro, para que ele não seja do conhecimento dos que detêm o poder de mando.

É preciso inverter a estrutura, colocando o cliente como a pessoa mais importante da organização. Mas isso não pode ser apenas um discurso de boas intenções. Vai exigir mudanças para as quais existem duas palavras-chave: delegação e cooperação; disseminação das informações.

1 SATISFAÇÃO DO CLIENTE

Colocar a pessoa certa, na hora certa, para fazer certo, da primeira vez, o que o cliente deseja. Este é o padrão de excelência desejado. Mas será que existe mesmo na organização o "lugar certo" para essa "pessoa certa"? Será que a estrutura da empresa está direcionada à satisfação total do cliente?

A maioria das empresas possui estruturas tradicionais de comando, onde o cliente relaciona-se com as pessoas que têm menor poder de decisão. Existe uma distância enorme entre os que detêm o poder (a direção superior) e o cliente.

O ovo de Colombo é revirar totalmente esta estrutura superada. E adotar a "pirâmide invertida" da Qualidade Total. Aí, as pessoas mais importantes na organização passam a ser as de atendimento e vendas.

São elas que têm contato direto com o cliente. Os demais funcionários são responsáveis pelo bom desempenho do pessoal de frente. A direção fica na base da pirâmide: seu papel é dar sustentação à finalidade de bem atender.

A delegação de poder é fundamental nesse tipo de organização participativa e cooperativa. Os acontecimentos mais importantes não são as reuniões de chefia, mas os momentos em que a empresa tem contato direto com o cliente, os "momentos da verdade".

A empresa passa a estruturar-se para transformar em sucesso esses "momentos da verdade". Por isso, os clientes estão no topo do organograma. Todos os demais setores se transformam em fornecedores de facilidade para os eventos de satisfação do cliente. O fluxo de operações estará direcionado para o atendimento do cliente.

2 RELACIONAMENTO COOPERATIVO

A responsabilidade compartilhada e o trabalho em equipe só poderão se desenvolver se a estrutura permitir uma interação constante entre as áreas.

A empresa toda é um macroprocesso, uma equipe única voltada para o objetivo comum de atingir altos níveis de produtividade, com a manutenção e a conquista de novos clientes.

3 DISSEMINAÇÃO DE INFORMAÇÕES

O fluxo de informações que parte do cliente (pedidos, avaliações, reclamações, expectativas) passa pelos diferentes departamentos da empresa e deve retornar como resposta e solução, de maneira ágil, ao cliente. O fluxo da decisão e a cadeia cliente-fornecedor devem estar alinhados aos valores que o cliente preza: cortesia, presteza, eficiência, receptividade e personalização.

É conhecido o fenômeno de "ruído na comunicação": a informação se enfraquece e deforma quanto maior o número de transmissores e receptores intermediários. É comum a gerência desconhecer a realidade da operação "na ponta".

A perda de competitividade é a consequência mais direta da falta de agilidade nas decisões, perda de informações, aumento da burocracia interna, pois, nos "momentos da verdade", o funcionário precisa tomar decisões que implicam, muitas vezes, questões vitais para o cliente. Frequentes consultas aos níveis superiores causam perda de tempo e dinheiro. A redução dos níveis hierárquicos ao mínimo necessário traz agilidade. Experimente.

Sumários

Referência*: A satisfação total do cliente é a meta fundamental de uma empresa moderna. O alcance desse objetivo significa o sucesso dela. Nas empresas tradicionais, o culto da figura do chefe ou diretor levava os funcionários a não corrigir erros, mas a ocultá-los a fim de não desagradar-lhe. Na empresa moderna, é primordial inverter essa primazia. O que vai exigir delegar decisões, incentivar o espírito de cooperação e disseminar informações. Os funcionários que contatam diretamente com o cliente devem ter bastante autonomia de decisões, a fim de que este confie neles. Portanto, o seu nível de competência para tal função deve ser o melhor possível. Os demais participantes da empresa devem constituir-se em suporte para a sua eficiência. O relacionamento cooperativo deve ser resultado da interação estrutural entre as áreas. Ele pressupõe compartilhar responsabilidades, evitar estrelismos, com vistas a otimizar a produção e a conquistar novos clientes. O pedido de informações ou sugestões dadas pelo cliente devem fluir rapidamente pelos diversos departamentos da empresa, e o resultado retornar com a mesma presteza, para que as decisões do funcionário-atendente conquistem o cliente e não o percam para o concorrente.*

Baseline-0: *É preciso inverter as estruturas para diminuir a distância entre o cliente e os que detêm o poder de decisão. Empresa fadada ao insucesso tem duas caras: uma real, outra para o cliente. Na hora de vender promessas quando o cliente confere decepções. É difícil encontrar o responsável quando a empresa não é direcionada à satisfação total dos clientes. Nas estruturas tradicionais de empresas onde o mando predomina sobre a responsabilidade individual não é possível sequer aprender em cima dos próprios erros. Faz se de tudo para que não haja registro do erro para que ele não seja do conhecimento dos que detêm o poder de mando. É preciso inverter a estrutura, colocando o cliente como a pessoa mais importante da organização. Mas isso não pode ser apenas um discurso de boas intenções. Vai exigir mudanças para as quais existem duas palavras-chave: delegação e cooperação; disseminação das informações. SATISFAÇÃO DO CLIENTE. Colocar a pessoa certa, na hora certa, para fazer certo, da primeira vez, o que o cliente deseja. Este é o padrão de excelência desejado. Mas será que existe mesmo na organização o lugar certo para essa pessoa certa?*

Baseline-1 e Baseline-2: *É preciso inverter as estruturas para diminuir a distância entre o cliente e os que detêm o poder de decisão. Empresa fadada ao insucesso tem duas caras: uma real, outra para o cliente. Na hora de vender promessas quando o cliente confere decepções. É difícil encontrar o responsável quando a empresa não é direcionada à satisfação total dos clientes. A empresa toda é um macroprocesso, uma equipe única voltada para o objetivo comum de atingir altos níveis de produtividade, com a manutenção e a conquista de novos clientes. O fluxo de informações que parte do cliente (pedidos, avaliações, reclamações, expectativas) passa pelos diferentes departamentos da empresa e deve retornar como resposta e solução, de maneira ágil ao cliente. O fluxo da decisão e a cadeia cliente fornecedor devem estar alinhados aos valores que o cliente preza: cortesia, presteza, eficiência, receptividade e personalização. A perda de competitividade é a consequência mais direta da falta de agilidade nas decisões, perda de informações,*

aumento da burocracia interna, pois, nos momentos da verdade, o funcionário precisa tomar decisões que implicam, muitas vezes, questões vitais para o cliente. A redução dos níveis hierárquicos ao mínimo necessário traz agilidade. Experimente.

GistSumm: *É difícil encontrar o responsável, quando a empresa não é direcionada à satisfação total dos clientes. É preciso inverter a estrutura, colocando o cliente como a pessoa mais importante da organização. A maioria das empresas possui estruturas tradicionais de comando, onde o cliente relaciona-se com as pessoas que têm menor poder de decisão. Os acontecimentos mais importantes não são as reuniões de chefia, mas os momentos em que a empresa tem contato direto com o cliente, os "momentos da verdade". A empresa toda é um macroprocesso, uma equipe única voltada para o objetivo comum de atingir altos níveis de produtividade, com a manutenção e a conquista de novos clientes. O fluxo de informações que parte do cliente (pedidos, avaliações, reclamações, expectativas) passa pelos diferentes departamentos da empresa e deve retornar como resposta e solução, de maneira ágil, ao cliente. A perda de competitividade é a consequência mais direta da falta de agilidade nas decisões, perda de informações, aumento da burocracia interna, pois, nos "momentos da verdade", o funcionário precisa tomar decisões que implicam, muitas vezes, questões vitais para o cliente.*

FSumm: *Empresa fadada ao insucesso tem duas caras: uma real, outra para o cliente. Na hora de vender, promessas; quando o cliente confere, decepções. É difícil encontrar o responsável, quando a empresa não é direcionada à satisfação total dos clientes. Será que a estrutura da empresa está direcionada à satisfação total do cliente? A empresa passa a estruturar-se para transformar em sucesso esses momentos da verdade. A empresa toda é um macroprocesso, uma equipe única voltada para o objetivo comum de atingir altos níveis de produtividade, com a manutenção e a conquista de novos clientes. O fluxo de informações que parte do cliente (pedidos, avaliações, reclamações, expectativas) passa pelos diferentes departamentos da empresa e deve retornar como resposta e solução, de maneira ágil, ao cliente. O fluxo da decisão e a cadeia cliente fornecedor devem estar alinhados aos valores que o cliente preza: cortesia, presteza, eficiência, receptividade e personalização. A perda de competitividade é a consequência mais direta da falta de agilidade nas decisões, perda de informações, aumento da burocracia interna, pois, nos "momentos da verdade", o funcionário precisa tomar decisões que implicam, muitas vezes, questões vitais para o cliente.*